ESSAYS ON THE ECONOMETRICS OF CAUSAL INFERENCE

By

Qi Xu

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Economics

August 11th, 2023

Nashville, Tennessee

Approved:

Atsushi Inoue, Ph.D.

Tong Li, Ph.D.

Tatsushi Oka, Ph.D.

Pedro H.C. Sant'Anna, Ph.D.

ACKNOWLEDGMENTS

I want to express my gratitude to numerous people who have made this dissertation possible. Foremost among these are my esteemed advisors, Tong Li and Atsushi Inoue. They didn't just introduce me to the world of econometric theory but also stood by me as amazing life mentors. Their encouragement, guidance, and their knack for sparking new ideas have been instrumental to my progress. They've always been open to discussing their thoughts and quickly responding to mine. Without their relentless effort and assistance, the successful completion of this dissertation would have remained a distant reality.

In addition, I wish to express my deep gratitude towards my coauthor and my first econometric professor at Vanderbilt, Pedro Sant'Anna. His passion for econometrics and commitment to academic excellence have always inspired me. It has been an absolute pleasure to have had the opportunity to collaborate with him on multiple ventures.

A special thanks goes out to Tatsushi Oka, with whom I've had numerous productive discussions while working on our joint project. He taught me a great deal about innovative ideas, advanced techniques, and the importance of meticulousness.

My journey at Vanderbilt was also greatly enriched by the friendship and support of my friends - Cameron, Elijah, Jun, Martin, Oscar, Rachel, Thu, Trang, and Zizheng. From trivia nights to hiking adventures and board games, their companionship added tremendous joy to my journey.

Last but not least, my heartfelt appreciation goes to my family, especially my parents. Their unwavering support, even from afar, has been a constant source of strength for me. They've made all my endeavors possible with their love and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Sensitivity Analysis for Treatment Effects with Endogenously Censored Duration Outcome

1.1 Introduction

Many program evaluation problems involve censored outcomes. A few classical examples include survival time of patients in clinical trials, duration of unemployment, length of marriage, the lifetimes of firms, and so forth. In addition to the usual problems with counterfactual analyses, censoring poses additional challenges to researchers, as it is well known that the marginal and joint distributions of the latent outcome and the censoring variable are not identifiable if the censoring mechanism is left entirely unrestricted (Tsiatis, 1975). One popular approach to restoring identification is by assuming that the two variables are independent, possibly conditional on observed covariates. Although prevalent, such an assumption can easily be violated in many practically relevant applications. Subject attrition due to unobserved factors being correlated with the latent outcome, and the presence of competing events are among the most frequently encountered reasons for the failure of the proposed assumption. As a concrete example, consider clinical patients who receive poor prognosis. They may decide to withdraw from trials based on such result, causing a positive correlation between survival and abandonment times. Ignoring this dependence would lead to biased assessments of the treatment.

Methodologies under dependent censoring have received less attention relative to their independent counterpart, partly due to a lack of consensus on how the dependence should be modeled. The current literature is divided between imposing a known censoring mechanism and making no assumptions about it at all. At one extreme, if we assume that the dependence structure is fully characterized by a known copula, we can recover distributional information of the latent duration from observed durations (see Zheng and Klein (1995)). However, such results are sensitive to the specification of the true copula. On the other extreme, robust approaches, such as the ones proposed by Khan and Tamer (2009) and Khan et al. (2016), utilize minimal theoretical restrictions and are likely to generate uninformative identified sets.

In reality, researchers often have some prior information on the censoring mechanism, which may come from auxiliary data, scientific theory, or expert opinion. For instance, in the clinical trial example, we may assume the latent and censoring times are positively correlated based on prior research findings. It is crucial that identification and inference procedures built by researchers allow one to flexibly incorporate partial information as such when addressing policy relevant questions.

With this goal in mind, we follow the partial identification approach proposed by Fan and Liu (2018), and assume that the true copula of latent outcome and censoring time belongs to the well-known Archimedean family. We do not, however, directly specify the true copula. The Archimedean copulas serve two purposes here. First, it allows the distribution of potential outcome be explicitly expressed in the form of copula-graphic-type estimands (see Rivest and Wells (2001)) that we denominate as *bound generating functions* (BGF). Such functions are smooth functionals of the observed (sub-) distributions and are indexed by the level of dependence censoring. Second, many one-parameter Archimedean families are endowed with a concordance ordering (Nelsen, 2007). The BGFs inherit such a property, and as a result, are ordered in terms of the *first-order stochastic dominance* (FOSD) relations. This natural ordering then allow us to explicitly derive the bounds of identified sets for various treatment effects. To the best of our knowledge, this copula-based partial identification approach to program evaluation has not yet been covered in the literature.

Analytical forms of the bounds render sensitivity analysis with respect to the level of dependent censoring especially convenient. With endogenously censored data, such analyses are crucial for obtaining convincing policy assessments, because the assumptions on the censoring mechanism are intrinsically untestable.

Our second contribution is the incorporation of the single-index structure into the aforementioned copula-based approach, based on which, we propose estimation procedures for the BGFs, as well as the bounds of the treatment effects, using a novel *single-index copula graphic* (SICG) estimator. The dimension-reduction feature of this new estimator is particularly attractive in the current context, where fully nonparametric methods such as those adopted by Braekers and Veraverbeke (2005), Lopez (2011), and Fan and Liu (2018), tend to be plagued by the "curse of dimensionality", due to the multitude of baseline covariates needed for justifying the unconfoundedness setting.

We provide comprehensive large sample results for the proposed estimators, including a uniform linear expansion for the new SICG estimator. Based on these, we establish functional central limit theorems for the BGFs as well as the bounds of the treatment effects. To conduct uniformly valid inference, we propose easy-to-implement multiplier bootstrap procedures, and show the bootstrap uniform confidence sets are asymptotically accurate.

We illustrate the proposed methodology through Monte Carlo studies and an empirical application in which we compare the relative efficacy of two treatment protocols for acute lymphoblastic leukaemia (ALL): GHS-2000 and AHOPCA ALL-2008. Using data from a series of clinical studies conducted in Honduras, prior work by Bernasconi et al. (2022) found that the more recent treatment plan leads to better survival prospects for patients in the first three years post treatment. Their results depend crucially on the potential survival time and abandonment-of-treatment being independent conditionally on observed covariates. When we depart from this assumption, however, this conclusion may not continue to hold according to the results of our sensitivity analysis.

Related literature: This article contributes to an extensive literature on program evaluation with censored data. The majority of works in this literature rely on the random censoring assumption. See, e.g. Anstrom and Tsiatis (2001), Hubbard et al. (2000), Lee and Lee (2005), Frandsen (2015), Sant'Anna (2016), Sant'Anna (2021), and so on. Models that accommodate dependent censoring are gaining attention. For instance, Beyhum et al. (2021), and Crommen et al. (2022) both study inference problems with endogenous treatment models. Our paper differs from these two, as we neither impose strong completeness nor functional-form assumptions on the data generating process (DGP), and we do not aim for point identification.

This article is also related to the literature on dependent censoring and competing risk models. Early contribution by Tsiatis (1975) shows that the joint distribution of the competing risks is not identified, and the best obtainable bounds are the worst-case bounds derived by Peterson (1976). These results allude to the difficulty of accounting for endogenous censoring without extra constraints or external information. To model the dependence between the potential outcome and the censoring variable, we follow the copula-based approach. With a fully known copula, Zheng and Klein (1995) propose a nonparametric estimator, which extends the one by Kaplan and Meier (1958), and call it the copula-graphic estimator. Rivest and Wells (2001) show that the estimator has a closed-form expression when attention is restricted to the Archimedean copulas. Braekers and Veraverbeke (2005), Huang and Zhang (2008), and Chen (2010) further extend it by incorporating covariates. The known copula assumption is imposed in all of the above works. In a linear quantile regression setting, Fan and Liu (2018) propose a partial identification approach that allows copula to vary within a prespecified class. This paper extends their approach to the program evaluation framework. More recently, Czado and Van Keilegom (2021), Deresa and Van Keilegom (2020), and Deresa et al. (2022) also allow an unknown copula, but they establish its identifiability via strong distributional restrictions.

We also build on the literature of single-index estimation with censored data. As a powerful dimension-reduction device, single-index models are widely popular in semiparametric duration analysis, cf. Lopez (2011), Lopez et al. (2013), and Bouaziz and Lopez (2010). The available results all rely on the random censoring assumption, and are not directly applicable to the copula-based setting. The novel SICG estimator proposed in this paper fills this gap. It is worthwhile to mention that it is not restricted to the scope of this article and can be used in many other settings. Additionally, since we follow Li and Patilea (2018) and impose the single-index structure directly on the potential laws, rather than on the observed distributions, our estimation procedure for the index parameter is greatly simplified relative to the aforementioned papers.

Organization of the article: Section 1.2 introduces the framework for endogenous censoring and the treatment effect parameters. Section 1.3 presents the single-index model, and in addition, introduce Archimedean copulas, and the bound generating functions. We also provide identification results on the aforementioned quantities in this section. Next, in Section 1.4, we propose a multi-step estimation procedure for various treatment effects, using the identification results derived in Section 1.3. We also establish large sample theories for the proposed estimators in this section. Section 1.5 establishes the validity of multiplier bootstrap procedures, and provide practical guidelines for constructing uniform bootstrap confidence bands. In Section 1.6, we illustrate the finite sample performance of our proposed estimators and the bootstrap confidence sets, via Monte Carlo simulations. Section 1.7 presents an empirical application, and Section 1.8 concludes. Proofs and auxiliary results are collected in Section 1.9.

1.2 Setup and Parameter of Interest

1.2.1 Model Framework

Consider a program in which the outcome of interest is measured by the amount of time until a target event occurs. Let $T \in \mathscr{T} \subset [0,\infty)$ denote such an outcome. We also observe an indicator D for binary treatment: $D = 1$ if the unit is treated and $D = 0$ otherwise. Following Neyman-Rubin potential outcome framework (see e.g. Rubin (1974)), we denote by T_1 and T_0 the values that T would have taken if D is equal to one or zero, respectively. As a result, $T = DT_1 + (1-D)T_0$. A vector $X \in \mathscr{X} \subset \mathbb{R}^k$ of baseline covariates is recorded prior to the program. In an ideal setting, we would observe (T_1, T_0, D, X), and make inferences thereof. However, the ideal data is coarsened in two ways.

First, we only observe the realized event time T but not the potential outcomes T_1 and T_0. Moreover, the realized T is subject to right censoring by a random variable, $C \in \mathscr{T}$. As a result, we only have access to $Y = \min\{T, C\}$ and a no-censoring indicator R, where $R = 1$ if $T \leq C$, and $R = 0$, otherwise. Same as the outcome of interest, C, Y and R are also functions of D, i.e. $U = DU_1 + (1-D)U_0$, where $U \in \{C, Y, R\}$ and U_d stands for the potential realization of U under treatment d. Thus, the available data W consists of (Y, R, D, X). Let $S_{U|V}(\cdot|v) = \mathbb{P}(U > \cdot|V = v)$ denote the *survival function* of a random variable U given $V = v$. Our goal is to make inferences on functionals of $S_{T_d|X}$, using information from observed samples of W.

In observational studies, treatment is not randomly assigned. Therefore, the treatment, the event time, and the censoring variable are all likely confounded. To address the relationship between the treatment and latent duration outcomes, we focus on the unconfoundedness setup. That is, we impose the following assumption on the underlying data generating process,

Assumption 1.1 (Unconfoundedness) $(T_1, T_0, C_1, C_0) \perp\!\!\!\perp D|X$.

Assumption 1.1 implies that, selection into treatment is solely based on observable characteristics. The assumption is akin to the standard unconfoundedness condition in the program evaluation literature, as found in complete observations (cf. Rosenbaum and Rubin (1983), Hirano et al. (2003), and Firpo (2007)), as well as censored outcomes (cf. Lee and Lee (2005), and Sant'Anna (2016, 2021)). It differs from the latter two, however, by requiring independence of the joint law, rather than on the potential event time only. This strengthened condition is necessary since the event and censoring times can remain dependent, even after adjusting for the covariates.

Since the observed duration Y and the censoring indicator R are deterministic functions of T and C, the above assumption immediately implies that $(Y_1, Y_0, R_1, R_0) \perp\!\!\!\perp D|X$. We also note that since the experimental setting can be viewed as a special case of Assumption 1.1, all of our theories presented below will automatically carry over to the randomized-controlled-trial setting.

1.2.2 Parameters of Interest

We will work mainly with the following four types of treatment effects under the unconfoundedness setup:

$$\begin{aligned}
&\text{(Restricted) Average Treatment Effect: } ATE(t) \equiv \mathbb{E}\left[\tilde{T}_1(t) - \tilde{T}_0(t)\right],\\
&\text{Distributional Treatment Effect: } DTE(t) \equiv F_{T_1}(t) - F_{T_0}(t),\\
&\text{Quantile Treatment Effect: } QTE(\tau) \equiv F_{T_1}^{-1}(\tau) - F_{T_0}^{-1}(\tau),\\
&\text{Cumulative Hazard Treatment Effect: } CHTE(t) \equiv \Lambda_{T_1}(t) - \Lambda_{T_0}(t),
\end{aligned}$$

where $\tilde{T}_d(t) \equiv \mathbb{1}\{T_d \leq t\} \cdot T_d + \mathbb{1}\{T_d > t\} \cdot t$ is generated from T_d by censoring the latter at t. The restricted ATE converges to the usual ATE as t increases. We adopt this restricted measure over the global one primarily because of the challenges posed by the right tail of the potential outcome distributions. As we will discuss in Section 1.4, they become increasingly close to boundaries as t grows, and therefore, increasingly difficult to estimate accurately, which in turn causes inference issues for the global ATE measure.[1] This measure is also used by Westling et al. (2021).

The quantile function $F_{T_d}^{-1}(\cdot)$ and the cumulative hazard function $\Lambda_{T_d}(\cdot)$ associated with treatment type $d \in \{0,1\}$ are defined by $F_{T_d}^{-1}(\tau) \equiv \inf\{y : F_{T_d(y)} \geq \tau\}$ and $\Lambda_{T_d} : F_{T_d} \mapsto \int_{[0,\cdot]} \frac{1}{1-F_{T_d}^{-}} dF_{T_d}$, with $F^{-}(x) \equiv \lim_{s \uparrow t} F(s)$, respectively. Note that each of these policy effects can be represented as the difference between smooth functionals of F_{T_d}. These functionals, denoted by $F_{T_d} \mapsto \Upsilon(F_{T_d}(\cdot))(\cdot)$, are usually called the *treatment responses*. It can be shown that each of the treatment responses introduced here respects the FOSD relations of F_{T_d}. That is, either $\Upsilon(F(\cdot))(u) \geq \Upsilon(G(\cdot))(u)$ or $\Upsilon(G(\cdot))(u) \geq \Upsilon(F(\cdot))(u)$ for all u, whenever $F(t) \geq G(t)$, for all t. Such a property will be exploited for characterizing identified set for the treatment effects. There are examples of treatment responses that violate the FOSD relation. For instance, the Gini coefficients and Lorenz curves respect second-order stochastic dominance relations but not the first-order one. Consequently, our identification analysis do not apply in these cases.

Under independence censoring mechanism, the policy parameters are known to be point identified from the observed data. See e.g. Hubbard et al. (2000), Lee and Lee (2005), and Sant'Anna (2016). However, when the censoring

[1] Another censored ATE measure, frequently encountered in the literature, is defined as $ATE(t) = \mathbb{E}[T_1 \mathbb{1}\{T_1 \leq t\} - T_0 \mathbb{1}\{T_0 \leq t\}]$. However, since the treatment response $\mathbb{E}[T_d \mathbb{1}\{T_d \leq t\}]$ does not respect the first order stochastic dominance relations of F_{T_d}, it is incompatible with the analytical framework adopted in this paper.

mechanism is entirely unrestricted, the best attainable result is the worst-case bounds by Peterson (1976). We aim to take the middle ground in this paper, and try to address the following type of question: if the level of dependence can be restricted to a given range, what values of the treatment effects are consistent with this information? The answer depends on two factors: (i) the quantification of the level of dependence censoring and (ii) a link from the censoring mechanism to the policy parameters. These two ingredients are discussed in detail in the next section.

1.3 Identification

We describe our identification strategy in this section. Our main result can be divided into two parts. We first introduce a single-index model, and discuss the identification of its index parameters. Then, in the second part, we provide the identification results on the distributions of potential durations and the treatment effects through the lens of copula theory. Despite the order of our exposition, the majority of results in the second part are relatively independent, and can be established without the embedding of the single-index structure.

1.3.1 Single-Index Model

Semiparametric models offer a good compromise between the parametric approach, which relies on strong assumptions on the functional form assumption that may not hold in practice, and the fully nonparametric one, which suffers from the curse of dimensionality. A well-known example of such a dimension reduction device is the single-index model, widely adopted in duration analysis. See, for instance, Xia et al. (2010), Bouaziz and Lopez (2010), Lopez et al. (2013), Li and Patilea (2018), and Bücher et al. (2021). For a generic conditional distribution of Y given X, the single-index model assumes that $F_{Y|X}(y|x) = G(y, x'\gamma^{\dagger})$, where G is an unknown bivariate function, and $\gamma^{\dagger}$ is the vector of index parameters. In general, the coefficients are only identified up to a scale, thus requiring normalization for point identification. To this end, we arrange the covariates such that the first k_1 variables are absolutely continuously distributed, and the remaining k_2 variables are binary. We set the coefficient associated with the first element, $x_{[1]}$, to 1, and let $x\gamma = x_{[1]} + x'_{[-1]}\gamma$, where $x_{[-1]}$ collects all the other covariates and γ is the corresponding subvector of $\gamma^{\dagger}$. Note that this normalization is not entirely innocuous as it imposes a positive effect on the first component of the covariates.

Assumption 1.2 (Single index structure) $(T_d, C_d) \perp\!\!\!\perp X | X\gamma_d$, where γ_d is an interior point of a compact set $\Gamma \subset \mathbb{R}^{k-1}$, for $d \in \{0,1\}$.

Assumption 1.2 is an index sufficiency condition on the joint law of the event time and the censoring variable. A similar restriction appears in Li and Patilea (2018) under the random censoring mechanism. Note that the true index coefficient may vary across treatment groups, reflecting potential differences in the treatment response heterogeneity. However, the indices are restricted to be the same across the marginal laws of T_d and C_d, for each $d \in \{0,1\}$. An immediate consequence of the index sufficiency condition is that $X\gamma_d$ can be viewed as a balancing score, meaning that the potential outcomes are independent of the treatment choices conditional on this index. The result is formally stated in Lemma 1.1.

Lemma 1.1 Under Assumptions 1.1 and 1.2, $(T_d, C_d, Y_d, R_d) \perp\!\!\!\perp D|X\gamma_d$, for $d \in \{0,1\}$.

The lemma essentially states that the property of unconfoundedness, as induced by the conditioning set X, is maintained under a coarse partition of X generated by the index $X\gamma_d$. This matching condition is crucial for establishing the identification of $\boldsymbol{\gamma} \equiv (\gamma_1', \gamma_0')'$ from the observed data.

With slight abuse of notation, we write $G_{d,r}(\cdot, x\gamma) = F_{Y,R|D,X\gamma}(\cdot, r|d, x\gamma)$, and define $f_d(u) = \partial F_{X\gamma,D}(u,d)/\partial u$, for $(d,r) \in \{0,1\}^2$, where the functional form of $G_{d,r}$ and f_d depends on γ. Furthermore, we define

$$\mathscr{E}_{d,r,\gamma}(t) \equiv \mathbb{1}\{D=d\}\left\{\mathbb{1}\{R=r, Y \le t\} - G_{d,r}(t, X\gamma)\right\}$$
$$U_{d,\gamma}(t,d,r) \equiv \mathscr{E}_{d,r,\gamma}(t) f_d(X\gamma),$$

and let $\mathscr{E}_{d,r,\gamma,\ell}$ and $U_{d,\gamma,\ell}$ be the same functions defined with observation W_ℓ.

Exploiting the balancing property of the index, we will show in Theorem 1.1 that, under the index sufficiency condition, $\mathbb{E}[U_{d,\gamma_d,\ell}(t,r)|X] = 0$ almost surely, for each t, d, and r. This conditional moment restriction will serve as the basis for the identification of the index parameters. To fully exploit the informational content of such a conditional restriction, we will follow the "integrated conditional moment approach" common in the specification testing literature. See, e.g. González-Manteiga and Crujeiras (2013) for a review. The idea is to characterize the conditional moment restriction as an infinite number of unconditional moment equations via some well-chosen family of weight functions $\{\vartheta(X;z) : z \in \mathscr{Z}\}$. That is,

$$\mathbb{E}[U_{d,\gamma_d,1}(t,r)|X] = 0 \text{ a.s.} \Leftrightarrow \mathbb{E}[U_{d,\gamma_d}(t,r)\vartheta(X;z)] = 0 \text{ a.e. in } z \in \mathscr{Z}, \tag{1.3.1}$$

Lemma 1 of Escanciano (2006b) provides primitive conditions on the family of weights for the equivalence in the preceding display to hold. Here, we list a few examples that satisfy the equivalence condition: (i) $\vartheta(X;z) = \mathbb{1}\{X \le z\}$, with $z \in \mathbb{R}^k$, see e.g., Stute (1997) and Domínguez and Lobato (2004); (ii) $\vartheta(X;z) = \mathbb{1}\{X'z_1 \le z_2\}$, with $z = (z_1, z_2) \in \mathbb{S} \times \mathbb{R}$, where $\mathbb{S}^k$ is the $(k-1)$-dimensional unit sphere, see e.g. Escanciano (2006a); (iii) $\vartheta(X;z) = \exp(iz'X)$, with $z \in \mathbb{R}^k$ and $i = \sqrt{-1}$, see e.g., Bierens (1982) and Lavergne and Patilea (2013).

Now, we define

$$\mathscr{J}_d(\gamma;\vartheta) \equiv \int_{\mathscr{T}\times\{0,1\}} \int_{z\in\mathscr{X}} \left\|\mathbb{E}[U_{d,\gamma}(t,r)\vartheta(X;z)]\right\|^2 d\Pi_Z(z) d\Pi_{T,R}(t,r), \tag{1.3.2}$$

where Π_Z is an integrating measure that is absolutely continuous with respect to the dominant measure of z. Similarly, $\Pi_{T,R}$ is an integrating measure for (t,r) that is specified by the researcher. It is not necessarily related to the unobserved law of T. In Theorem 1.1, we will show that $\gamma_d = \arg\min_{\gamma\in\Gamma} \mathscr{J}_d(\gamma;\vartheta)$, $d\in\{0,1\}$, and the minimization yields a unique solution, if the conditions given in Assumption 1.3 are fulfilled.

Assumption 1.3 (Identification of index)

1. (i) $\mathscr{X} = \mathscr{X}^c \times \mathscr{X}^b \equiv \Pi_{\ell_1=1}^{k_1}[\underline{x}_{\ell_1}, \bar{x}_{\ell_1}] \times \{0,1\}^{k_2}$; (ii) $\inf_{x\gamma\in\mathscr{X}_\Gamma} f_d(X\gamma) > 0$, where $\mathscr{X}_\Gamma \equiv \{x\gamma : x\in\mathscr{X}, \gamma\in\Gamma\}$; (ii) $\mathscr{T} = [0,\bar{y}]$, where $\bar{y} = \inf\{y : \inf_{(r,d,x)\in\{0,1\}^2\times\mathscr{X}} F_{Y_d,R_d|X\gamma_d}(y,r|x\gamma_d) = 1\}$.
2. $\mathbb{P}(D=d|X) > 0$, almost surely.
3. There exist sets $\mathscr{T}_0 \subset \mathscr{T}$, such that for each $t\in\mathscr{T}_0$, (i) the function $z \mapsto F_{Y_d,R_d|X\gamma_d}(t,r|z)$ is differentiable in z; (ii) there exists a set $\mathscr{X}_0 \subset \mathscr{X}$, such that $\mathbb{P}(X\in\mathscr{X}_0) > 0$, and $\partial F_{Y_d,R_d|X\gamma_d}(t,r|v)/\partial v|_{v=x\gamma_d} \neq 0$, for all $x\in\mathscr{X}_0$.
4. For each $\gamma\in\Gamma$, there exists an open interval $\mathscr{V}_0$ satisfying (i) $\mathscr{V}_0 \subset \cap_{\ell=0}^{k_2-1}\{\mathscr{X}_\gamma^c + \gamma_{k_1+\ell}\} \cap \mathscr{X}_\gamma^c$, where $\mathscr{X}_\gamma^c = \{x_{[1]} + \gamma_1 x_{[2]} + \ldots + \gamma_{k_1-1}x_{[k_1]} : (x_{[1]},\ldots,x_{[k_1]}) \in \mathscr{X}^c\}$, and (ii) for each $t\in\mathscr{T}_0$, if $F_{Y_d,R_d|X\gamma_d}(t,r|v+u) = F_{Y_d,R_d|X\gamma_d}$ $(t,r|v)$ for all $v\in\mathscr{V}_0$, then $u=0$.

Assumptions 1.3.1 and 1.3.2 are standard. We allow for continuous covariates as well as discrete ones. Here, the discrete variables are all assumed to be binary, but the restriction can be easily relaxed. Note that Assumption 1.2.1(ii) implies that $\bar{t} \leq \bar{c}$, where $\bar{t}$ and $\bar{c}$ are the upper bounds in the support of the event time and the censoring variable, respectively. Outcome beyond $\bar{y}$ will never be observed, thus the entire distribution $F_{T_d|X\gamma_d}$, and thus the ATE, can be identified only if $\bar{t} \leq \bar{c}$. When the interest lies in functionals that do not involve the entire distribution, this assumption is not needed. Assumption 1.3.2 is the usual overlapping condition on the treatment assignment mechanism, imposed to guarantee that the conditional distribution $G_{d,r}$ are well defined on $\mathscr{X}_\Gamma$. The next two conditions, adapted from Assumptions 4.1 and 4.2 in Ichimura (1993), are imposed to ensure the identifiability of the index parameters. Together with the normalization restriction, Assumption 1.3.3 secures identification of index coefficients corresponding to the continuous covariates. Assumption 1.3.4 restricts the shape of $\mathscr{X}_\Gamma$, and when it is assumed in addition, the coefficients for binary covariates are point identified as well.

Theorem 1.1 Under Assumptions 1.1, 1.2, 1.3.1, and 1.3.2 it holds that (i)

$$\mathbb{E}[U_{d,\gamma_d}(t,r)|X]=0, \text{ almost surely}, \forall(d,r,t)\in\{0,1\}^2\times\mathscr{T}. \tag{1.3.3}$$

(ii) If in addition, Assumptions 1.3.3 and 1.3.4 hold, $\gamma\neq\gamma_d$ implies $\mathbb{E}[U_{d,\gamma}(t,r)|X]\neq 0$, almost surely, for all $(d,r,t)\in\{0,1\}^2\times\mathscr{T}_0$. (iii) If in addition, ϑ belongs to any of the classes of functions in Lemma 1 of Escanciano (2006b), and $\int_{\mathscr{T}_0\times\{0,1\}}d\Pi_{T,R}(t,r)>0$, we have that $\mathscr{J}_d(\gamma;\vartheta)\geq 0$, $\forall\gamma\in\Gamma$, and the equality holds if and only if $\gamma=\gamma_d$, for $d\in\{0,1\}$.

Theorem 1.3 is a global identification result. It shows that the index parameters can be recovered as the unique minimizer of the minimum distance type criterion, (1.3.2). Compared with similar approaches by Bouaziz and Lopez (2010), Strzalkowska-Kominiak and Cao (2014), and Li and Patilea (2018), we do not directly impose the uniqueness of single-index structure, but rather derive it from primitive and mild conditions on the underlying DGP.

As an implication of Theorem 1.3, we show how $\boldsymbol{\gamma}$ can be estimated based on a reformulation of $\mathscr{J}_d(\gamma;\vartheta)$. Towards this end, we note that, by means of the law of iterated expectations, $\mathscr{J}_d(\gamma,\vartheta)$ can be written as

$$\mathscr{J}_d(\gamma;\rho)=\int_{\mathscr{T}\times\{0,1\}}\mathbb{E}\left[\rho(X_1,X_2)U_{d,\gamma,1}(t,r)U_{d,\gamma,2}(t,r)\right]d\Pi_{T,R}(t,r), \tag{1.3.4}$$

where $\rho(x_1,x_2)\equiv\int_{z\in\mathscr{Z}}\vartheta(x_1,z)^c\vartheta(x_2,z)d\Pi_Z(z)$, and A^c is the conjugate transpose of A. The function ρ might appear complicated at first, but a convenient closed-form usually follows once an appropriate weight function and integrating measure combination is chosen. For instance, when $\vartheta(X;z)=\exp(iz'X)$ and $\Pi_Z(z)=\Phi(z)$, where $\Phi(\cdot)$ is the CDF of k-variate standard normal distribution, $\rho(x_1,x_2)=\exp(-\|x_1-x_2\|/2)$.[2] Other examples can be found in Escanciano (2006b) and Sant'Anna et al. (2022). The new criterion (1.3.4) follows the minimum distance function of Li and Patilea (2018) closely, inheriting many attractive properties of theirs. For one, since the dimension reduction hypothesis is imposed on joint laws of T_d and C_d, (1.3.2) does not involve complicated conditional Kaplan-Meier type integrals as is common in the literature. See, e.g. Xia et al. (2010), Bouaziz and Lopez (2010), and Strzalkowska-Kominiak and Cao (2014). Moreover, no trimming is required, due to the inclusion of f_d in $U_{d,\gamma}$. As such, we may avoid dealing with convoluted multi-step procedures as appeared in Delecroix et al. (2006), and Bouaziz and Lopez (2010).

We propose to estimate the index parameters $\boldsymbol{\gamma}$ by minimizing the sample analogue of (1.3.4). The integration with

[2] Here we have used the fact that

$$\int_{\mathbb{R}^k}\exp(iu't)\cdot\frac{\exp(-u'u/2)}{(2\pi)^{k/2}}du=\mathbb{E}_U[\exp(iU't)]=\exp(-t't/2),$$

where the first equality follows by the definition of characteristic function for random variable U, and the second is due to the assumption that U follows the k-variate standard normal distribution.

respect to $\Pi_{T,R}$ may also be avoided when the it is replaced by a suitable empirical measure. Details of the estimation procedure are provided in Section 1.4.

Remark 1.1 The index sufficiency condition can be tested based on (1.3.4), following the approach proposed by Maistre and Patilea (2019). Let $Q(\gamma)$ be a $(k-1)\times(k-1)$-invertible matrix with the first column given by γ. Consider the following function

$$\mathscr{J}_d(\gamma;\rho,g) \equiv \int_{\bar{\mathscr{T}}\times\{0,1\}} \mathbb{E}\left[\rho(X_1'Q(\gamma),X_2'Q(\gamma))U_{d,\gamma,1}(t,r)U_{d,\gamma,2}(t,r)J_g(X_2\gamma,X_1\gamma)\right] d\Pi_{T,R}(t,r),$$

where $J(\cdot)$ is a symmetric kernel function, g is a bandwidth, and $J_g(u,v)\equiv g^{-1}J(g^{-1}(v-u))$. If Assumption 1.2 does not hold, after adjusting for $X\gamma$, the dependence on X would drive $\mathscr{J}_d(\gamma;\rho,g)$ away from zero, uniformly in $\gamma\in\Gamma$ and for a suitably chosen bandwidth sequence. In turn, we may construct the test statistic based on sample analogues of $\mathscr{J}_d(\gamma,\rho,g)$, and use multiplier bootstrap to generate the critical values. The idea is formalized in Algorithm 1.9.2.

Remark 1.2 We would like to emphasize here that point identification of the index parameters does not imply that of the marginal distribution of T_d, as well as the joint distribution of T_d and C_d. For the latter, Peterson (1976)'s worse case bounds are applicable here and they are equivalent to the fully nonparametric case, provided that Assumption 1.2 indeed holds.

1.3.2 Partial Identification through Copula

From Sklar (1959)'s theorem, we know that, conditionally on $X=x$, there exists a conditional survival copula, $\mathscr{C}_x(\cdot,\cdot):[0,1]^2\mapsto[0,1]$, such that

$$\mathbb{P}\left(T_d>t,C_d>c|X=x\right)=\mathscr{C}_x(S_{T_d|X}(t|x),S_{C_d|X}(c|x)),$$

for $t,c\in\mathscr{T}$. Moreover, if the conditional survival functions are absolutely continuous, then $\mathscr{C}_x$ is unique; otherwise it is only uniquely determined on the range of the survival functions. Sklar's results allow us to separate the analysis of the marginal laws and the dependence structure. As is discussed in Section 1.1, we mainly focus on parameter Archimedean families, but similar identification result can be established for nonparametric families as well. Introduced by Genest and MacKay (1986a; 1986b), Archimedean copulas are widely used in economic applications for modeling a variety of dependence structures. The family is characterized by a generator function $\phi_\theta(\cdot):[0,1]\mapsto[0,\infty)$ that is usually indexed by a parameter $\theta\in\Theta$:

$$\left\{\mathscr{C}(u,v;\theta)=\phi_\theta^{[-1]}\left(\phi_\theta(u)+\phi_\theta(v)\right):\theta\in\Theta\right\}.$$

For each θ, ϕ_θ is a known continuous, convex, strictly decreasing function with $\phi_\theta(1) = 0$. In the above definition, $\phi_\theta^{[-1]}$ stands for the pseudo-inverse of ϕ_θ, as defined by

$$\phi_\theta^{[-1]}(s) = \begin{cases} \phi_\theta^{-1}(s), & 0 \leq s \leq \phi_\theta(0) \\ 0 & \phi_\theta(0) \leq s \leq \infty. \end{cases}$$

If $\phi_\theta(0) = \infty$, $\phi_\theta^{[-1]} = \phi_\theta^{-1}$, and the copula is said to be *strict*.

We do not seek to identify parameter θ in this article. Instead, we treat it as a sensitivity parameter that is varied, to trace out a family of identified sets. Prior information on the dependence structure, such as model restrictions and expert opinions, will be translated as constraints on θ, which will serve to restrict the size of the identified set.

In place of a random censoring condition, the censoring mechanism of this paper are defined through mild restrictions on the copula functions as seen below.

Assumption 1.4 (Copula)

1. (i) The conditional distribution of (T_1, T_0, C_1, C_0) is absolutely continuous respect to the Lebesgue measure. (ii) Conditional distributions $F_{Y_d,R_d|D,X}(y,r|d,x)$ and $F_{T_d|X}(y|x)$ are differentiable with respect to $y \in \mathscr{T}$, for $x \in \mathscr{X}$ and $r,d \in \{0,1\}$.
2. The true conditional survival copula of (T_d, C_d), $\mathscr{C}_x(\cdot,\cdot)$, is strict and belongs to the one parameter Archimedean family $\mathscr{C}(\cdot,\cdot;\theta)$ with generator function $\phi_\theta(\cdot)$ indexed by $\theta \in \Theta \equiv [\underline{\theta}, \bar{\theta}]$, for any $x \in \mathscr{X}$ and $d \in \{0,1\}$.
3. Let $\phi_\theta^{'}(u) = \partial \phi_\theta(u)/\partial u$, for $u \in (0,1)$. It holds that $\phi_{\theta_1}^{'}(\cdot)/\phi_{\theta_2}^{'}(\cdot)$ is strictly increasing for any $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$.

Assumption 1.4.1 is a smoothness condition on the duration outcomes, requiring that the event and censoring times admit densities on the support of Y. Discretely-measured times pose an additional challenge for identification. Accommodating discrete outcomes along the lines of Kim (2021) will be left for future work. Assumption 1.4.2 stipulates that the conditional copulas belong to an Archimedean family with the same type of generator function, but the index parameters are potentially different depending on x. While the assumption streamlines the discussion of identification and sensitivity analysis, it is not restrictive and can be relaxed to allow for nonparametric generator functions. We let $\theta_d^*(x)$ denote the true copula parameter associated with $\mathscr{C}_x$ under treatment d. Archimedean copulas can be justified in the context of a mixed proportional hazards model with common frailty term, but it is not restricted to such models. See e.g. Joe (1997) and Nelsen (2007) for detailed expositions. Many Archimedean families include the independence copula, $\phi_\theta(u) = \log u^{-1}$, either as a special case or as a limiting one. This feature is particularly convenient for sensitivity analysis using Archimedean copulas. Assumption 1.4.3 and Corollary 4.4.6 in Nelsen (2007) imply the family of copulas is endowed with a *concordance* ordering, meaning that $\mathscr{C}(u,v;\theta_1) \leq \mathscr{C}(u,v;\theta) \leq \mathscr{C}(u,v;\theta_2)$, for all $u,v \in [0,1]^2$ and $\theta \in [\theta_1,\theta_2]$. As a result, θ sufficiently characterizes the level of dependence between T_d and C_d. This property plays a major part in generating analytical bounds for the treatment effects.

Generator functions satisfying Assumption 1.4.3 are common. A few well-known examples include: (i) *Clayton* copula: $\max\{u^{-\theta}+v^{-\theta}-1,0\}^{-1/\theta}$, with the generator $\phi_\theta(u)=\frac{1}{\theta}(u^{-\theta}-1)$; (ii) *Gumbel* copula: $\exp(-[(-\log u)^\theta + (-\log v)^\theta]^{1/\theta})$, with the generator $\phi_\theta(u)=(\log u^{-1})^\theta$; (iii) *Gumbel-Hougaard* copula: $uv\exp(-\theta\log u\log v)$, with the generator $\phi_\theta(u)=\log(1-\theta\log u)$. For a comprehensive list, see Table 4.1 in Nelsen (2007) and Table 1 in Fan and Liu (2018).

Provided that the true copula belongs to the Archimedean family, the distribution of T_d can be explicitly expressed, in terms of the generator function ϕ_θ, $G_{d,1}$, and $s_d \equiv 1-G_d$, $d\in\{0,1\}$. We denote the linking function as the conditional *bound generating function* (BGF), which is defined as follows

$$s_{T_d}(t,x\gamma,\theta) \equiv \phi_\theta^{-1}\left(-\int_0^t \phi_\theta'\left(s_d(y,x\gamma)\right)G_{d,1}(dy,x\gamma)\right), \tag{1.3.5}$$

for $(t,d,x\gamma,\theta)\in\mathscr{T}\times\{0,1\}\times\mathscr{X}_\Gamma\times\Theta$. For a function $\theta(\cdot):\mathscr{X}\to\Theta$, the unconditional BGF, $s_{T_d,\gamma}(\cdot,\theta(\cdot)):\mathscr{T}\times\mathscr{X}_\Gamma\to[0,1]$, is defined by $s_{T_d,\gamma}(t,\theta(\cdot))=\mathbb{E}[s_{T_d}(t,X\gamma,\theta(X))]$. We suppress the subscript γ, when it is evaluated at its true value.

The next theorem formally states how the conditional and marginal distributions of T_0 and T_1 can be recovered from the observed (conditional) distributions, via the BGFs.

Theorem 1.2 For $d\in\{0,1\}$, (i) under Assumptions 1.1–1.3, 1.4.1, and 1.4.2, $S_{T_d|X}(\cdot|x)\in\left\{s_{T_d}(\cdot,x\gamma_d,\theta):\theta\in\Theta\right\}$, a.e. for $x\in\mathscr{X}$, and $S_{T_d}(\cdot)\in\left\{s_{T_d}(\cdot,\theta(\cdot)):\theta(x)\in\Theta,x\in\mathscr{X}\right\}$. The identified sets are uniformly sharp over $\mathscr{T}$, for $d\in\{0,1\}$.

(ii) Suppose Assumptions 1.1–1.4 hold. If in addition, for any $\theta_1,\theta_2\in\Theta$ such that $\theta_1\leq\theta_d^*(\cdot)\leq\theta_2$, we have $s_{T_d}(t,x\gamma_d,\theta_2)\leq S_{T_d|X}(t|x)\leq s_{T_d}(t,x\gamma_d,\theta_1)$, and $s_{T_d}(t,\theta_2)\leq S_{T_d}(t)\leq s_{T_d}(t,\theta_1)$, a.e. for $(t,x)\in\mathscr{T}\times\mathscr{X}$. The identified sets are uniformly sharp across t and x, for each $d\in\{0,1\}$.

As a direct implication of the theorem, when the true copula is known, or equivalently when $\underline{\theta}=\bar{\theta}$, the distribution of potential event time can be point identified from data. The result thus extends Rivest and Wells (2001) and Braekers and Veraverbeke (2005)'s findings to the program evaluation framework. Furthermore, this implies that, when the outcome is randomly censored, the potential survival distribution can be recovered with the famous Kaplan-Meier estimator as proposed by Beran (1981) and Dabrowska (1989). We remark that when the index sufficiency condition fails, conclusions of Theorem 1.2 will continue to hold with (1.3.5) replaced by its nonparametric counterpart.

Remark 1.3 There is no way to learn about the true copula family from data. Prior work including Zheng and Klein (1995), Huang and Zhang (2008), Lo and Wilke (2010), and Fan and Liu (2018) finds that the choice of generating functions are less important than that of the level of dependence θ. Extensive numerical evidence suggests biases caused by misspecification of the copula family are negligible compared to that caused by θ. As such, the choice of copula family itself does not accord much identification power.

Remark 1.4 Due to the convexity of the generator function and by construction, the function $t \mapsto s_{T_d}(t, x\gamma_d, \theta)$ is monotonically decreasing and bounded between $[0,1]$ for each $(d,\theta) \in \{0,1\} \times \Theta$. These constraints may not be respected when s_{T_d} is replaced by its estimator. We discuss a remedy in remark 1.5.

The name BGF is motivated by the fact that the ordering of such functions, as induced by the concordance ordering of the copula, yields a convenient characterization for the bounds of the treatment effects. Each type of the treatment effects introduced in Section 1.2 consists of treatment responses that respect the FOSD relations of S_{T_d}. As a result, bounds of BGFs can be translated to those of the treatment responses. Exploiting this insight, we derive closed-form bounds for various treatment effects in the next proposition.

Let us denote $q_{d,\gamma,\theta}(\tau) \equiv \inf\{y : s_{T_d,\gamma}(y,\theta) \leq 1-\tau\}$ as the τ-th quantile of $1 - s_{T_d,\gamma}(\cdot,\theta)$, for $\tau \in (0,1)$. Again, the subscript γ is omitted when it is evaluated at its true value.

Proposition 1.1 Suppose that Assumptions 1.1–1.4 hold, and that $\theta_1 \leq \theta_d^*(\cdot) \leq \theta_2$, for $d \in \{0,1\}$ and $\theta_1, \theta_2 \in \Theta$. Then, we have

$$v_{ATE}(t,\boldsymbol{\theta}) \in \left[\int_{[0,t]} \left(s_{T_1}(y,\theta_2) - s_{T_0}(y,\theta_1)\right) dy,\ \int_{[0,t]} \left(s_{T_1}(y,\theta_1) - s_{T_0}(y,\theta_2)\right) dy\right], \tag{1.3.6}$$

$$v_{DTE}(t,\boldsymbol{\theta}) \in [s_{T_0}(t,\theta_2) - s_{T_1}(t,\theta_1),\ s_{T_0}(t,\theta_1) - s_{T_1}(t,\theta_2)], \tag{1.3.7}$$

$$v_{QTE}(\tau,\boldsymbol{\theta}) \in [q_{1,\theta_2}(\tau) - q_{0,\theta_1}(\tau),\ q_{1,\theta_1}(\tau) - q_{0,\theta_2}(\tau)], \tag{1.3.8}$$

$$v_{CHTE}(t,\boldsymbol{\theta}) \in [\log(s_{T_0}(t,\theta_2)) - \log(s_{T_1}(t,\theta_1)),\ \log(s_{T_0}(t,\theta_1)) - \log(s_{T_1}(t,\theta_2))], \tag{1.3.9}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $t \in \mathscr{T}$ and $\tau \in (0, \bar{\tau})$. The identified sets are uniformly sharp across t or τ, depending on j.

For a fixed $x \in \mathscr{X}$, bounds for the conditional treatment effects can be determined by replacing $s_{T_d}(t,\theta)$ and $q_{d,\theta}(\tau)$ in the previous displays with $s_{T_d}(t, x\gamma_d, \theta)$ and $q^x_{d,\theta}(\tau) \equiv \inf\{y : s_{T_d}(y, x\gamma_d, \theta) \leq 1-\tau\}$, respectively.

For each $j \in \{ATE, DTE, QTE, CHTE\}$, we let the lower and upper bound be denoted by $v_{lb,j}(u,\boldsymbol{\theta})$ and $v_{ub,j}(u, \boldsymbol{\theta})$, and let them by denominated as the lower and upper overall *treatment effect bound functions* (TEBF) for type j, respectively. We use the vector $\boldsymbol{\nu}_j = (\nu_{lb,j},\ \nu_{ub,j})'$ to collect the bounds.[3] With slight abuse of notation, the index variable u, is allowed to vary depending on the type of treatment effect under consideration. In particular, $u = \tau$ if $j = QTE$, and $u = t$ if $j \in \{ATE, DTE, CHTE\}$.

Proposition 1.1 is the first contribution of the paper. It implies that, if the true copula parameter mappings ($\theta_0^*(\cdot)$, $\theta_1^*(\cdot)$) lies between constants θ_1 and θ_2, the bounds of treatment effects can be expressed as smooth functionals of $\{(s_{T_d}(\cdot,\theta_1),\ s_{T_d}(\cdot,\theta_2))\}_{d\in\{0,1\}}$. Moreover, the lower and upper bounds are related by $\nu_{lb}(\cdot,\boldsymbol{\theta}) = \nu_{ub}(\cdot,\breve{\boldsymbol{\theta}})$, where $\breve{\boldsymbol{\theta}} = (\theta_2,\theta_1)$. The size of the identified set is determined by the strength of prior information. As the interval $[\theta_1,\theta_2]$ narrows, the identified sets become smaller. In the limit, the true copula is known, and the treatment effects can be point identified. This result generalizes those found in Lee and Lee (2005) and Sant'Anna (2016), to the case where independent censoring is no longer maintained. Again, even if the index sufficiency condition fails, results of Proposition 1.1 can be preserved with appropriate modifications to the BGFs.

1.4 Estimation and Large Sample Theory

The estimation of TEBFs consists of three steps. We sketch the steps here, in an informal way to illustrate the main idea, and the details are provided in subsequent subsections. In the first step, we estimate the index parameters $\boldsymbol{\gamma}$ by minimizing an estimator of (1.3.4). Next, we construct a consistent estimator for the conditional and average BGFs. For this purpose, we propose a new single-index copula graphic estimator. In the last step, we construct estimates of TEBFs using the BGFs estimated in the preceding step.

1.4.1 Single-Index Parameters

As we mentioned earlier, when $\Pi_{T,R}$ is chosen to be $F_{Y,R}$, the empirical analogue of (1.3.4) admits an analytical expression. Hence, we keep this choice fixed for the remainder of the paper. We can also use other integrating measures that only involve the observed laws. For instance, we may set $\Pi_{T,R} = F_{Y,R|D=d}$ or $\Pi_{T,R} = F_{Y,R=1|D=d}$, correspondingly for $U_{d,\gamma}$ and $d \in \{0,1\}$. Given a univariate kernel function $L(\cdot)$ and a bandwidth b that changes with sample size n, we define the sample analogue of U_γ and $\mathscr{J}$ as

$$\begin{aligned}
\hat{U}_{d,\gamma,i}(y,r) &= \frac{1}{n-1}\sum_{j=1}\left\{I_{d,y,r,i} - I_{d,y,r,j}\right\} L_b(X_i\gamma, X_j\gamma),\\
\hat{\mathscr{J}}_d(\gamma;\rho) &= \frac{1}{n^2}\sum_{\ell=1}^{n}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\rho(X_i,X_j)\left(\hat{U}_{d,\gamma,i}(Y_\ell,R_\ell)\hat{U}_{d,\gamma,j}(Y_\ell,R_\ell)\right)\right\},
\end{aligned} \tag{1.4.1}$$

[3] Analogous remarks apply to the conditional TEBFs, $\boldsymbol{\nu}_j^x = (\nu_{lb,j}^x, \nu_{ub,j}^x)'$, where the definitions for $\nu_{lb,j}^x$ and $\nu_{ub,j}^x$ should be apparent.

where $I_{d,y,r,\ell} = D_\ell \mathbb{1}\{R_\ell = r, Y_\ell \leq y\}$, and $L_b(x,y) = b^{-1} L\left(b^{-1}(y-x)\right)$. Then, for a user-specified weighting function $\rho(\cdot)$, we estimate γ_d by minimizing $\hat{\mathscr{J}}_d$. That is,

$$\hat{\gamma}_d = \arg\min_{\gamma \in \Gamma} \hat{\mathscr{J}}_d(\gamma; \rho), \tag{1.4.2}$$

for $d \in \{0,1\}$.

Under the regularity conditions to be specified in Section 1.4.3, we can show that the proposed index estimator is consistent, converges at the parametric rate, admits an asymptotic linear representation, and converges to a normal distribution. These results are established in Section 1.9.2. Among these results, it is worth noting that the consistency and convergence rate are particularly useful for establishing the uniform expansion and other properties of the conditional BGF estimator. Similar results for the unconditional case would further hinge on the existence of the linear representation.

1.4.2 BGF Estimators

Exploiting Proposition 1.2, estimators of BGFs can be constructed from estimators of the index coefficient $\boldsymbol{\gamma}$ and the observed distributions $\{G_{d,r}\}_{d,r \in \{0,1\}}$. For the latter, we propose to use the Nadaraya-Watson-type kernel estimator. Specifically, for any $\gamma \in \Gamma$, we let

$$\hat{G}_{d,r}(y, x\gamma) = \frac{\frac{1}{n}\sum_{i=1}^n I_{d,y,r,i} K_h(x\gamma, X_i\gamma)}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{D_i = d\} K_h(x\gamma, X_i\gamma)} \equiv \frac{\hat{\kappa}_{d,r,y}(x\gamma)}{\hat{f}_d(x\gamma)}, \tag{1.4.3}$$

$$\hat{G}_d(y, x\gamma) = \hat{G}_{d,1}(y, x\gamma) + \hat{G}_{d,0}(y, x\gamma), \tag{1.4.4}$$

where $K(\cdot)$ is a univariate kernel function, potentially different from $L(\cdot)$, and h is a bandwidth parameter.[4] The observed survival function estimator is given by $\hat{s}_d = 1 - \hat{G}_d,$. Replacing s_d and $G_{d,1}$ in (1.3.5) with these estimators, we get

$$\hat{s}_{T_d}(t, x\gamma, \theta) \equiv \phi_\theta^{-1}\left\{-\frac{1}{\hat{f}_d(x\gamma)n}\sum_{i=1}^n \phi'_\theta\left(\hat{s}_d(Y_i, x\gamma)\right) I_{d,t,r,i} K_h(x\gamma, X_i\gamma)\right\}, \tag{1.4.5}$$

[4] It is well known that kernel estimators exhibit large bias around the boundary points. In practice, we may modify estimators of $\hat{\kappa}_{d,r,y}$ and $\hat{f}_d$ as follows to avoid the boundary issue

$$\tilde{\kappa}_{d,r,y}(x\gamma) = \hat{\kappa}_{d,r,y}(z), \text{ and } \tilde{f}_d(x\gamma) = \hat{f}_d(z),$$

where $z = \min\{\mathscr{X}_\gamma\} + h$ if $x\gamma \in [\min\{\mathscr{X}_\gamma\}, \min\{\mathscr{X}_\gamma\} + h]$, $z = \max\{\mathscr{X}_\gamma\} - h$ if $x\gamma \in [\max\{\mathscr{X}_\gamma\} - h, \max\{\mathscr{X}_\gamma\}]$; otherwise, $z = x\gamma$. We keep the untransformed estimator to simplify technical analysis. Results can be extended to the modified estimators with relative ease.

and $\hat{s}_{T_d}(t,\theta) = n^{-1}\sum_{i=1}^n \hat{s}_{T_d}(t, X_i\hat{\gamma}_d, \theta)$, for all $t, d, x,$ and θ. We note that (1.4.5) is an extension of the plug-in estimator of Fan and Liu (2018) to single-index models. Although their estimator can also accommodate multivariate (continuous) covariates, it is, however, not generally recommended due to the curse of dimensionality. We call $\hat{s}_{T_d}$ the *single index copula graphic* (SICG) estimator, because it is first-order asymptotically equivalent to the following estimator

$$\tilde{s}_{T_d}(t, x\gamma, \theta) \equiv \phi_\theta^{-1}\left\{ -\sum_{Y_i \le t}^{n} R_i \left(\phi_\theta(\hat{s}_d(Y_i, x\gamma)) - \phi_\theta(\hat{s}_d(Y_i, x\gamma) - w_{i,n}(x,\gamma)) \right) \right\},$$

where $w_{i,n}(x,\gamma) \equiv \mathbb{1}\{D_i = d\} K_h(x\gamma, X_i\gamma)/(n\hat{f}_d(x\gamma))$. This estimator directly adapts the nonparametric copula graphic estimator in Braekers and Veraverbeke (2005) to the single-index models. Due to the first-order equivalence, inference results in Section 1.4 and 1.5 automatically apply to $\tilde{s}_{T,d}$ as well.

When the independence copula is assumed, i.e. $\phi_\theta(u) = -\log u$, (1.4.5) becomes

$$\hat{s}_{T_d,ind}(t, x\gamma) = \exp\left(-\sum_{i=1}^{n} \frac{w_{i,n}(x,\gamma) R_i}{\sum_{j=1}^n w_{j,n}(x,\gamma)\mathbb{1}\{Y_j > Y_i\}} \right) \tag{1.4.6}$$

$$\approx \prod_{Y_i \le t} \left(1 - \frac{w_{i,n}(x,\gamma)}{\sum_{j=1}^n w_{j,n}(x,\gamma)\mathbb{1}\{Y_j > Y_i\}} \right)^{R_i}, \tag{1.4.7}$$

where the asymptotic equivalence follows roughly from Taylor expanding the exponential function. When there is no treatment, (1.4.7) coincides with the conditional single-index Kaplan-Meier estimator, as proposed by Strzalkowska-Kominiak and Cao (2014) or Li and Patilea (2018). On the other hand, under random censoring, the average BGF can be directly estimated using the standard Kaplan-Meier estimator of Kaplan and Meier (1958), without going through the conditioning step.

Remark 1.5 Note that the random function $t \mapsto \hat{s}_{T_d}(t, x\hat{\gamma}_d, \theta)$ necessarily lies between $[0,1]$ due to the range constraint on $\phi_\theta^{-1}(\cdot)$. However, as discussed in the previous section, the estimator may not be monotonically decreasing in finite samples, even though its population counterpart is indeed constrained to be. To enforce the shape constraint, we plan to rearrange the initial estimator using the procedure proposed by Chernozhukov et al. (2010). This will result in a proper conditional survival function. From Chernozhukov et al. (2010), we find that the initial and rearranged estimators are asymptotically equivalent if $s_{T_d}(t, x\gamma_d, \theta)$ is indeed monotone. For this reason, our focus will be on the asymptotic results for the initial estimator in the following sections.

1.4.3 Uniform Linear Expansion

In this section, we will provide a linear expansion for the conditional SICG estimator $\hat{s}_{T_d}(t, x\hat{\gamma}_d, \theta)$ that is valid uniformly across t, x, and θ, based on which, a uniform linear representation for the unconditional SICG estimator $\hat{s}_{T_d}(t, \theta)$ is also derived. These uniform representations are crucial for establishing results on weak convergence and bootstrap validity. Here, we first introduce and discuss several assumptions necessary for deriving the claimed results.

Assumption 1.5 (Data)

1. The data $\{Y_i, R_i, D_i, X_i\}_{i=1}^n$ are independently and identically distributed.

2. There exists $y_o \in (0, \bar{y}]$ and $\upsilon_o > 0$ such that, for each $d \in \{0,1\}$, $F_{Y_d, R_d|X\gamma_d}(y_o, 1|x\gamma_d) \leq 1 - \upsilon_o$ almost surely in $x \in \mathscr{X}$. Let $\tilde{\mathscr{T}} \equiv [0, y_o]$.

Assumption 1.5.1 is standard. Assumption 1.5.2, which is also imposed by Rivest and Wells (2001), and Fan and Liu (2018), strengthens Assumption 1.2.1 by further restricting the support of the event time. Since many generator functions are not finite at 0, the condition is imposed to avoid dealing with a divergent $\phi^{-1}(\cdot)$, in a neighborhood of the origin.

For the following set of assumptions, we define a shrinking neighborhood of γ_d by $\Gamma_{d,n} \equiv \{\|\gamma - \gamma_d\| \leq Cn^{-1/2}\}$, for some positive constant C.

Assumption 1.6 (Smoothness)

1. For a positive integer $s \geq 2$ and $d = 0,1$, (i) The function $v \mapsto f_{d,\gamma_d}(v)$, is $(s+1)$-times continuously differentiable, and the derivatives up to order s are bounded; (ii) $\partial_v^{(s+1)} f_{d,\gamma_d}(v)$ is Lipschitz continuous in v with the Lipschitz constant being independent of v.

2. For $d, r \in \{0,1\}$, (i) the function $v \mapsto F_{Y,R|D,X\gamma_d}(y, r|d, v)$ is $(s+1)$-times continuously differentiable and the derivatives up to order s are bounded uniformly on $\tilde{\mathscr{T}}$; (ii) $\partial_v^{(s+1)} F_{Y,R|D,X\gamma_d}(y, r|d, v)$ is Lipschitz continuous in v with the Lipschitz constant being independent of y and v; (iii) $y \mapsto F_{Y,R|D,X\gamma_d}(y, r|d, v)$ is continuously differentiable and the first-order derivative is uniformly bounded with respect to v; (iv) $\partial_y F_{Y,R|D,X\gamma_d}(y, r|d, v)$ is Lipschitz continuous in both y and v, where the Lipschitz constants are independent of y, and v; (v) $\partial_v F_{Y,R|D,X\gamma_d}(y, r|d, v)$ is Lipschitz continuous in y with the Lipschitz constant being independent of v and y;

3. (i) The functions $v \mapsto \mathbb{E}[X_{[\ell]}|X\gamma_d = v]$, and $v \mapsto \mathbb{E}[X_{[\ell]}X_{[j]}|X\gamma_d = v]$, $\ell, j = 2, ..., k$, are four times continuously differentiable, and the derivatives up to fourth order are all bounded; (ii) the fourth order derivatives are Lipschitz continuous in v. For $r = 0,1$, $\ell_1 = 0,1$, and $\ell_2 = 0,1$, (iii) $v \mapsto \rho^{\gamma}_{\ell_1,\ell_2}(y, v)$ is continuously differentiable and the

derivatives are bounded uniformly on $\tilde{\mathscr{T}} \times \mathscr{X} \times \Gamma_{d,n}$;[5] (iv) $\partial_v \rho^{\gamma}_{\ell_1,\ell_2}(y,v)$ is Lipschitz continuous in v with the Lipschitz constant being independent of y, x, and $\gamma \in \Gamma_{d,n}$.

4. (i) $u \mapsto \phi_\theta(u)$ is three times continuously differentiable with the third order derivative $\phi^{'''}_\theta(u) \leq 0$ and $\phi^{'''}_\theta(u)$ being bounded uniformly for $(u,\theta) \in [\upsilon_o, 1] \times \Theta$; (ii) $1/\dot{\phi}^{-1}_\theta(z)$ and $\ddot{\phi}^{-1}_\theta(z)$ are bounded away from 0 for $(z,\theta) \in [0, y^*_o] \times \Theta$, where $\dot{\phi}^{-1}_\theta(z) \equiv \phi^{'}_\theta(\phi^{-1}_\theta(z))$, $\ddot{\phi}^{-1}_\theta(z) \equiv \phi^{''}_\theta(\phi^{-1}_\theta(z))$, and $y^*_o = (1-\upsilon_o)\sup_{(u,\theta)\in[\upsilon_o,1]\times\Theta} \left|\phi^{'}_\theta(u)\right|$; (iii) $\phi^{'}_\theta(u)$ and $\phi^{''}_\theta(u)$ are Lipschitz continuous in θ with Lipschitz constant being independent of $u \in [\upsilon_o, 1]$.

Assumption 1.6 gathers a set of smoothness conditions on various functions. Assumptions 1.6.1 and 1.6.2 are assumed in most of prior works, including Delecroix et al. (2006), Bouaziz and Lopez (2010), Xia et al. (2010), and Chiang and Huang (2012). Assumption 1.6.3 serves to bound the bias and to control the rate of first-order remainder terms. Assumption 1.6.4 stipulates that the generator functions exhibit enough smoothness with respect to both u and θ. These requirements, akin to Assumption (C8) in Braekers and Veraverbeke (2005) and Assumption G in Fan and Liu (2018), are necessary when establishing uniformity of the linear expansion of the SICG estimator with respect to θ. Now, we define $\psi^a_{d,r}(t,x) \equiv -\left(x_{[-1]} - \mathbb{E}[X_{[-1]}|x\gamma_d]\right) \partial G_{d,r}(t,v)/\partial v|_{v=x\gamma_d}$, and $V_d(t,r) \equiv \mathbb{E}\left[\psi^a_{d,r}(t,X)\psi^a_{d,r}(t,X)' f_d(X\gamma_d)^2\right]$.

Assumption 1.7 (Index Estimation)

1. (i) The class of functions $\{v \mapsto g_{d,r,\gamma}(v;t) : (d,r,t,\gamma) \in \{0,1\}^2 \times \mathscr{T} \times \Gamma\}$ is of the VC type with bounded envelop function,[6] where $g_{d,r,\gamma}(v;t)$ is either of the following functions and their derivatives up to the second order:

$$v \mapsto f_{d,\gamma}(v), \quad v \mapsto F_{Y,R|D,X\gamma}(t,r|d,v), \quad v \mapsto \mathbb{E}[X_{[\ell]}|X\gamma = v], \quad v \mapsto \mathbb{E}[X_{[\ell]}X_{[j]}|X\gamma = v],$$

for $\ell, j = 2, ..., k$; (ii) for $d \in \{0,1\}$ and each sequence $\delta_n \to 0$,

$$\sup_{\|\gamma-\gamma_d\|\leq\delta_n} \sup_{(t,r,v)\in\mathscr{T}\times\{0,1\}\times\mathscr{X}_\Gamma} \left|g_{d,r,\gamma}(v;t) - g_{d,r,\gamma}(v;t)\right| \to 0.$$

2. There exists a set $\mathscr{T}_v \subset \mathscr{T}$, such that $\mathbb{P}((Y,R) \in \mathscr{T}_v \times \{0,1\}) > 0$ and $V_d(t,r)$ is positive definite for each $(t,r,d) \in \mathscr{T} \times \{0,1\}^2$.

This assumption collects several regularity conditions needed for showing asymptotic behavior of the index estimator $\hat{\boldsymbol{\gamma}}$. The first condition provides uniform control for the local difference of second order derivatives of (1.4.1),

[5] For a matrix X, $X^{\otimes\ell}$ with $\ell = 0,1,2$ denote $1, X$, and XX', respectively. Define

$$\rho^{\gamma}_{\ell_1,\ell_2}(y,x\gamma) \equiv \partial^{\ell_2}_{x\gamma}\left\{f_d(x\gamma)\,\mathbb{E}\left[G^{\ell_1}_d(y,X\gamma_d)(x_{[-1]} - X_{[-1]})^{\otimes\ell_2}|X\gamma = x\gamma\right]\right\}. \tag{1.4.8}$$

[6] Precise definition of VC (Vapnik-Cěrvonenkis) type class is recalled in Section 1.9.3

while the second condition is imposed to guarantee that the Hessian matrix is positive definite, and thus, the asymptotic variance matrix is invertible.

Assumption 1.8 (Kernel)

1. The kernel function, $L(\cdot)$ is symmetric, supported on $[-1,1]$, and of bounded variation; (ii) it is twice continuously differentiable on $(-1,1)$ and the derivatives up to the second order are continuous and of bounded variation.

2. The kernel function, $K(\cdot)$ is symmetric, supported on $[-1,1]$ and of bounded variation; (ii) it is twice continuously differentiable on $(-1,1)$ with uniformly continuous and bounded derivatives; (iii) $\int K(u)du = 1$, $\int u^{\ell} K(u)du = 0$ for nonnegative integers $\ell < s$, and $\int u^s K(u)du < \infty$.

Assumption 1.9 (Bandwidth)

1. The bandwidth b satisfies: $b \to 0$, $\log n / \left(nb^3\right) \to \infty$, $nb^4 \to 0$, as $n \to \infty$.

2. The bandwidth h satisfies: $h \to 0$, $\log n/(nh^3) \to 0$, and $nh^{2s} \to 0$, as $n \to \infty$.

The restrictions on the kernel and the bandwidth are relatively mild. Assumptions 1.8.1 and 1.9.1 are imposed to ensure the estimation error from estimating $\boldsymbol{\gamma}$ is of the order less than $n^{-1/2}$. Meanwhile, Assumptions 1.8.2 and 1.9.2 provide rates control for the conditional SICG estimator. The smoothness conditions on the kernel functions serve two purposes: (1) it guarantees that $L(b^{-1}(\cdot\gamma - x\gamma))$ and $K(h^{-1}(\cdot\gamma - x\gamma))$ belong to the VC type class, which is necessary for establishing uniform convergence of several U-processes arising from the expansion of the kernel estimators; (2) it also allows us to control the rate of bias terms by means of Taylor expansions. Assumption 1.8 is satisfied by frequently used kernel functions, such as uniform, triangular, biweight, triweight, Epanechnikov kernels, etc. The Gaussian kernel, however, is ruled out due to the compact support condition.[7]

Due to the single-index structure, bandwidth conditions are independent of the dimension of X, k, meaning our estimator is not subject to the "curse of dimensionality". As a result, higher order kernels are not necessary when the covariate is multivariate. We require $nb^3/\log n, nh^3/\log n$ diverge to infinity, so that the first-order expansion of the kernel function with respect to γ is uniformly convergent. By imposing $nh^{2s} \to 0$, we undersmooth to make the bias disappear asymptotically.

Remark 1.6 As is the case for all semiparametric estimators, the smoothing parameters play a crucial role in the trade-off between reducing bias and variance. It is therefore desirable to have a data-adaptive way of choosing the

[7] Compactness of the kernel function is not essential, and can be relaxed by imposing conditions on the tail diminishing rate of the kernels. See, e.g. Maistre and Patilea (2019) for a detailed treatment.

parameter. One possibility is to estimate b and γ_d simultaneously via minimizing (1.4.1) with respect to (b,γ). In practice, we may follow a simple grid search procedure: (i) pick a finite grid $\{b_\ell\}_{\ell=1}^m$ from the set $[b_l n^{-\iota}, b_u n^{-\iota}]$, for some positive constants $b_l < b_u$ and some ι that fulfills Assumption 1.9.1. (ii) Minimize (1.4.1) with respect to γ, and record the minimum $\{\hat{\mathscr{J}}_d(b)\}_{d\in\{0,1\}}$ for each b in the grid, and keep the value of bandwidth such that $\{\hat{\mathscr{J}}_d(b)\}_{d\in\{0,1\}}$ attains the minimum value. When a second-order kernel is adopted, i.e. $s=2$, we may set h equal to b.

Now, we define a few more quantities related to the influence functions. Let $\mathscr{E}_{d,\gamma} = \sum_{r=0,1}\mathscr{E}_{d,r,\gamma}$, and $\psi_d^a = \sum_{r=0,1}\psi_{d,r}^a$. Moreover, $\psi_d^b(W) \equiv \int_{\tilde{\mathscr{T}}\times\{0,1\}} \mathbb{E}[\psi_{d,r}^a(y,X_1)f_d(X_1\gamma_d)\rho(X_1,X)|X]\ U_{d,\gamma_d}(y,r)dF_{Y,R}(y,r)$, and

$$V_d \equiv \int_{\tilde{\mathscr{T}}\times\{0,1\}} \mathbb{E}\left[\psi_d^a(y,X_1)\psi_d^a(y,X_2)' f_{d,\gamma_d}(X_1\gamma_d) f_{d,\gamma_d}(X_2\gamma_d)\rho(X_1,X_2)\right] dF_{Y,R}(y,r),$$

$$\Psi_d(f_1,f_2)(t,x,\theta) \equiv \frac{1}{\phi_\theta'\left(s_{T_d}(t,x\gamma_d,\theta)\right)}\left\{-\int_0^t \phi_\theta''\left(s_d(y,x\gamma_d)\right)f_1(y)G_{d,1}(dy,x\gamma_d)\right.$$
$$\left.+\phi_\theta'\left(s_d(t,x\gamma_d)\right)f_2(t) - \int_0^t \phi_\theta''\left(s_d(y,x\gamma_d)\right)f_2(y)s_d(dy,x\gamma_d)\right\},$$

where $\Psi_d(\cdot,\cdot)$ is a functional mapping from $\ell_\infty(\tilde{\mathscr{T}})\times\ell_\infty(\tilde{\mathscr{T}})$ to $\ell_\infty(\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta)$, for $d\in\{0,1\}$.[8]

Theorem 1.3 (Uniform asymptotic linear representation) Suppose Assumptions 1.1–1.9 hold,

$$\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta) - s_{T_d}(t,x\gamma_d,\theta) = \frac{1}{n}\sum_{i=1}^n \sum_{j\in\{s,b,l\}} \eta_{j,d}(W_i,x,t,\theta) + r_n(x,t,\theta)$$

where

$$\eta_{s,d}(W,x,t,\theta) = K_h(x\gamma_d, X\gamma_d)\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)/f_d(x\gamma_d),$$
$$\eta_{b,d}(W,x,t,\theta) = K_h(x\gamma_d, X\gamma_d)\Psi_d\left(G_d(\cdot,X\gamma_d)-G_d(\cdot,x\gamma_d), G_{d,1}(\cdot,X\gamma_d)-G_{d,1}(\cdot,x\gamma_d)\right)(t,x,\theta)/f_d(x\gamma_d),$$
$$\eta_{l,d}(W,x,t,\theta) = \psi_d^b(W)'V_d^{-1}\Psi_d\left(\psi_d^a,\psi_{d,1}^a\right)(t,x,\theta)/f_d(x\gamma_d)$$

and $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}|r_n(x,t,\theta)| = O_p\left((\log n)^{1/2}n^{-1}h^{-3/2}\right)$.

This theorem is the second main result of this article. It shows that, the conditional SICG estimator is asymptotically linear, and its influence functions can be split into four parts. The first two term, $\eta_{s,d}$ and $\eta_{b,d}$, are associated with the stochastic part and the bias of the usual kernel expansions. Of the two, the first component dominates in the limit with a uniform rate of $O_p\left((\log n)^{1/2}\cdot n^{-1/2}h^{-1/2}\right)$, free from the curse of dimensionality. The third component, $\eta_{l,d}$,

[8] For a generic set $\mathscr{S}$, $\ell_\infty(\mathscr{S})$ is the space of all uniformly bounded real functions on $\mathscr{S}$, equipped with the supremum norm, $\|f\|_{\mathscr{S}} \equiv \sup_{s\in\mathscr{S}}|f(s)|$.

is unique to the SICG estimator. It arises from the estimation of the index parameters and converges at the parametric rate, implying that the estimation error of the index coefficients is asymptotically negligible. Consequently, the main conclusions of the previous theorem remain intact when estimators other than $\hat{\boldsymbol{\gamma}}$ are used, provided that such estimators are root-n consistent.

1.4.4 Weak Convergence

This uniform linear representation allows us to apply techniques in the empirical process literature to establish weak convergence of the bound generating processes. The weak convergence, denoted by "$\Rightarrow$", is in the sense of Hoffmann–Jørgensen–Dudley, as recalled in Section 1.9.3. See, also Definition 1.3.3 of Van Der Vaart and Wellner (1996). The convergence takes place in $\ell_\infty(\mathscr{S})$.

Before stating the results, we will first need to introduce a few notations again. For $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta^2$, we collect the conditional and unconditional SICG estimators by $\hat{\mathbf{S}}^x(t, \boldsymbol{\theta}) \equiv \left(\hat{s}_{T_1}(t, x\hat{\gamma}_d, \theta_1),\ \hat{s}_{T_0}(t, x\hat{\gamma}_d, \theta_2)\right)'$, and by $\hat{\mathbf{S}}(t, \boldsymbol{\theta}) \equiv \left(\hat{s}_{T_1}(t, \theta_1),\ \hat{s}_{T_0}(t, \theta_2)\right)'$, respectively. Analogously, the BGFs are collected by $\mathbf{S}^x$ and $\mathbf{S}$, respectively.

We refer to $\hat{\mathbb{G}}_n^x(\cdot,\cdot) \equiv \sqrt{nh}\left(\hat{\mathbf{S}}^x(\cdot,\cdot) - \mathbf{S}^x(\cdot,\cdot)\right)$, for a fix $x \in \mathscr{X}$, as the *conditional bound generating process* (CBGP), and let $\hat{\mathbb{G}}_n(\cdot,\cdot) \equiv \sqrt{n}\left(\hat{\mathbf{S}}(\cdot,\cdot) - \mathbf{S}(\cdot,\cdot)\right)$ stand for the *unconditional bound generating process* (UBGP). The main goal of this section is to show that both CBGP and UBGP converge weakly to centered Gaussian processes. For this purpose, we need an additional assumption, which is given as follows.

Assumption 1.10 (i) $1/\dot{\phi}_\theta^{-1}(z)$ is Lipschitz continuous in θ with Lipschitz constant being independent of θ and $z \in [0, y_o^*]$; (ii) $u \mapsto \phi_\theta''(u)$ is $(s+1)$ times continuously differentiable with the $(s+1)$-th order derivative being bounded uniformly for $(u, \theta) \in [\upsilon_o, 1] \times \Theta$;

Assumption 1.10.(i) not only allows us to bound the derivative of $\phi_\theta^{-1}(\cdot)$ uniformly, but also contributes to controlling the size of the functional space associated with influence functions of the CBGP. Assumption 1.10.(ii) strengthens Assumption 1.6.4.(i). It ensures that the bias from approximating the UBGP by an empirical process is uniformly negligible. Most generator functions of the Archimedean family satisfy this stronger smoothness condition.

In the following corollary, we establish weak convergence of the CBGP. Its proof is the combination of Theorem 1.3 and Theorem 10.6 in Pollard (1990), the latter of which provides a set of sufficient conditions for the weak convergence of triangular arrays of non-identically distributed random elements.

Corollary 1.1 (i) Under the assumptions of Theorem 1.3, and suppose that Assumption 1.10.(i) hold, then $\hat{\mathbb{G}}_n^x(\cdot,\cdot) \Rightarrow \mathbb{G}^x(\cdot,\cdot)$, in $\ell_\infty(\tilde{\mathscr{T}} \times \Theta^2) \times \ell_\infty(\tilde{\mathscr{T}} \times \Theta^2)$, where $\mathbb{G}^x$ is a two-dimensional, tight, centered Gaussian process with covariance function,

$$\Sigma_\eta^x(\mathbf{t}, \boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{\varphi}^x(W, t_1, \boldsymbol{\theta}_1)\boldsymbol{\varphi}^x(W, t_2, \boldsymbol{\theta}_2)'\right],$$

for each $\mathbf{t} \equiv (t_1, t_2)' \in \tilde{\mathscr{T}} \times \tilde{\mathscr{T}}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')' \in \Theta^2 \times \Theta^2$, and for $j = 1, 2$, $\boldsymbol{\theta}_j = (\theta_j, \tilde{\theta}_j)$, and $\boldsymbol{\varphi}^x(w, t, \boldsymbol{\theta}_j) = (\eta_{s,1}(w, x, t, \theta_j), \eta_{s,0}(w, x, t, \tilde{\theta}_j))$.

To the best of our knowledge, this result is new to the literature. This result differs from Theorem 2 in Braekers and Veraverbeke (2005) in a number of ways. First, our CBGP is indexed not only by the time t but also by the copula parameter θ. In comparison, they study a similar process indexed by the time only. Consequently, our result generalizes theirs by relaxing the restrictive assumption that copula is completely known. Given the non-identifiability result from Tsiatis (1975), this generalization is a crucial initial step in our sensitivity analysis. Secondly, they consider a univariate fixed design for the covariates, whereas we adopt a single-index model that accommodates multivariate random variables, and allows them to be either discrete or continuous. Despite these differences, the covariance functions share a similar structure in the two papers. See Section 1.9.3.3 for formulas.

The following corollary records results on the UBGP that are parallel to Theorem 1.3 and Corollary 1.1.

Corollary 1.2 (i) Suppose the assumptions of Theorem 1.3, and Assumption 1.10 hold, we have that, for $(d, t, \theta) \in \{0, 1\} \times \tilde{\mathscr{T}} \times \Theta$,

$$\hat{s}_{T_d}(t, \theta) - s_{T_d}(t, \theta) = \frac{1}{n} \sum_{i=1}^{n} \varphi_d(W_i, t, \theta) + R_n(t, \theta),$$

where $\varphi_d = \sum_{j=1}^{2} \varphi_{d,j}$,

$$\begin{aligned}
\varphi_{d,1}(W, t, \theta) &= \Psi_d\left(\mathscr{E}_{d,\gamma_d}, \mathscr{E}_{d,1,\gamma_d}\right)(t, X, \theta) f(X\gamma_d) / f_d(X\gamma_d), \\
\varphi_{d,2}(W, t, \theta) &= s_{T_d}(t, X\gamma_d, \theta) - s_{T_d}(t, \theta),
\end{aligned}$$

and $\sup_{(t,\theta) \in \tilde{\mathscr{T}} \times \Theta} |R_n(t, \theta)| = o_p\left(n^{-1/2}\right)$.

(ii) Furthermore, $\hat{\mathbb{G}}_n(\cdot, \cdot) \Rightarrow \mathbb{G}(\cdot, \cdot)$, in $\ell_\infty(\tilde{\mathscr{T}} \times \Theta^2) \times \ell_\infty(\tilde{\mathscr{T}} \times \Theta^2)$, where $\mathbb{G}$ is a two-dimensional, tight, centered Gaussian process with covariance function

$$\Sigma_\varphi(\mathbf{t}, \boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{\varphi}(W, t_1, \boldsymbol{\theta}_1) \boldsymbol{\varphi}(W, t_2, \boldsymbol{\theta}_2)'\right],$$

for each $\mathbf{t} \equiv (t_1, t_2)' \in \tilde{\mathscr{T}} \times \tilde{\mathscr{T}}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')' \in \Theta^2 \times \Theta^2$, and for $j = 1, 2$, $\boldsymbol{\theta}_j = (\theta_j, \tilde{\theta}_j)$, and $\boldsymbol{\varphi}(W, t, \boldsymbol{\theta}_j) = (\varphi_1(W, t, \theta_j),\ \varphi_0(W, t, \tilde{\theta}_j))$.

As a first result, Corollary 1.2 provides a uniform linear expansion for the unconditional SICG estimator. The influence function can be decomposed into two parts. The first part, $\varphi_{d,1}$, comes from estimating the conditional

BGF. It is the linear representation of the first-order Hoeffding projection of the dominant U process. The second component, $\varphi_{d,2}$, arises from the sampling variation of X. Based on this uniform expansion, we show that the UBGP, as a process indexed by both t and θ, converges weakly to a centered Gaussian process. The rates of convergence are, however, different from the CBGP.

With these two corollaries in hand, we are equipped to present inference theories on the estimators of TEBFs. According to our discussion earlier, TEBFs are smooth functionals of the BGFs. It implies that, we can apply the functional delta method (see e.g. Theorem 3.9.5 in Van Der Vaart and Wellner (1996)), and show that the plug-in estimators of TEBFs will satisfy functional central limit theorems.

Now, let us define plug-in estimators for TEBFs. For $i = 1, ..., n$, denote the i-th order statistics of Y in the sample by $Y_{i:n}$. Let the quantile curve estimator be given by $\hat{q}_{d,\theta}(\tau) \equiv \inf\{y : \hat{s}_{T_d}(y, \theta) \leq 1 - \tau\}$, and its conditional version, by $\hat{q}^x_{d,\theta}(\tau) \equiv \inf\{y : \hat{s}_{T_d}(y, x\hat{\gamma}_d, \theta) \leq 1 - \tau\}$. With these notations, and in view of (1.3.6) - (1.3.9), we consider the following estimators of the lower TEBFs,

$$\hat{\nu}_{lb,ATE}(t, \boldsymbol{\theta}) \equiv \sum_{i=1}^{n-1} \mathbb{1}\left\{Y_{(i+1):n} \leq y_o\right\} \left(Y_{(i+1):n} - Y_{i:n}\right) \left(\hat{s}_{T_1}(Y_{i:n}, \theta_2) - \hat{s}_{T_0}(Y_{i:n}, \theta_1)\right), \tag{1.4.9}$$

$$\hat{\nu}_{lb,DTE}(t, \boldsymbol{\theta}) \equiv \hat{s}_{T_0}(t, \theta_2) - \hat{s}_{T_1}(t, \theta_1), \tag{1.4.10}$$

$$\hat{\nu}_{lb,QTE}(\tau, \boldsymbol{\theta}) \equiv \hat{q}_{1,\theta_2}(\tau) - \hat{q}_{0,\theta_1}(\tau), \tag{1.4.11}$$

$$\hat{\nu}_{lb,CHTE}(t, \boldsymbol{\theta}) \equiv \log(\hat{s}_{T,0}(t, \theta_2)) - \log(\hat{s}_{T,1}(t, \theta_1)), \tag{1.4.12}$$

where $\theta_1 \leq \theta_2$. $t \in \tilde{\mathscr{T}}$, and $\tau \in (0, \tau_o)$, where $\tau_o \equiv 1 - \sup_{(x,\theta) \in \mathscr{X} \times \Theta} s_{T_d}(y_o, x\gamma_d, \theta)$. Estimators of the upper TEBFs can be constructed by swapping the places of θ_1 and θ_2 on the right hand side of preceding equations. Here, we have restricted the upper bound of $\mathscr{T}$ to y_o, in order to avoid entering into the explosive tail area of the generator functions. Consistency of the ATE requires that y_o be sufficiently close to $\bar{y}$. Toward this end, we may set y_o as a large value close to $Y_{n:n}$ in practice.

To understand the formula for $\hat{\nu}_{lb,ATE}$, we note that $\hat{s}_{T_1}(t, \theta)$ and $\hat{s}_{T_0}(t, \theta)$ are step functions in t, with jumps at $\{Y_{i:n}\}_{i=1}^n$ only. The integral over $[0, t]$ can thus be divided into intervals with end points set by the order statistics. In each interval, the integrand is constant, yielding the product form.

In the remainder of this section, we will investigate the asymptotic behavior of $\sqrt{n}(\hat{\boldsymbol{\nu}}_j - \boldsymbol{\nu}_j)$ as well as $\sqrt{nh}(\hat{\boldsymbol{\nu}}^x_j - \boldsymbol{\nu}^x_j)$, for $j \in \{ATE, DTE, QTE, CHTE\}$. Again, let us introduce a few quantities, which are related to the influence functions of limiting processes. For $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta^2$, with $\theta_1 \leq \theta_2$, define $\boldsymbol{\psi}_j(W, u, \boldsymbol{\theta}) = (\psi_{1,j}(W, u, \theta_2) - \psi_{0,j}(W, u, \theta_1), \psi_{1,j}(W, u, \theta_1) - \psi_{0,j}(W, u, \theta_2))'$ and $\boldsymbol{\psi}^x_j(W, u, \boldsymbol{\theta}) = (\psi^x_{1,j}(W, u, \theta_2) - \psi^x_{0,j}(W, u, \theta_1), \psi^x_{1,j}(W, u, \theta_1) - \psi^x_{0,j}(W, u, \theta_2))'$, for $j \in \{ATE, QTE\}$. Meanwhile, we let $\boldsymbol{\psi}_j(W, u, \boldsymbol{\theta}) = (\psi_{1,j}(W, u, \theta_1) - \psi_{0,j}(W, u, \theta_2), \psi_{1,j}(W, u, \theta_2) -$

$\psi_{0,j}(W,u,\theta_1))'$ $\boldsymbol{\psi}^x_j(W,u,\boldsymbol{\theta}) = (\psi^x_{1,j}(W,u,\theta_1) - \psi^x_{0,j}(W,u,\theta_2), \psi^x_{1,j}(W,u,\theta_2) - \psi^x_{0,j}(W,u,\theta_1))'$, for $j \in \{DTE, CHTE\}$, where

$$\begin{aligned}
\psi_{d,ATE}(W,t,\theta) &= \int_{[0,t]} \varphi_d(W,y,\theta)dy, & \psi^x_{d,ATE}(W,t,\theta) &= \int_{[0,t]} \eta_{s,d}(W,x,y,\theta)dy,\\
\psi_{d,DTE}(W,t,\theta) &= -\varphi_d(W,t,\theta), & \psi^x_{d,DTE}(W,t,\theta) &= -\eta_{s,d}(W,x,t,\theta),\\
\psi_{d,QTE}(W,\tau,\theta) &= \frac{\varphi_d(W,q_{d,\theta}(\tau),\theta)}{f_{T_d}(q_{d,\theta}(\tau),\theta)}, & \psi^x_{d,QTE}(W,\tau,\theta) &= \frac{\eta_{s,d}(W,x,q^x_{d,\theta}(\tau),\theta)}{f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)},\\
\psi_{d,CHTE}(W,t,\theta) &= -\frac{\varphi_d(W,t,\theta)}{s_{T_d}(t,\theta)}, & \psi^x_{d,CHTE}(W,t,\theta) &= -\frac{\eta_{s,d}(W,x,t,\theta)}{s_{T_d}(t,x\gamma_d,\theta)},
\end{aligned}$$

for $d \in \{0,1\}$.

The next theorem establishes uniform central limit theorems for the conditional and overall TEBF estimators.

Theorem 1.4 (i) Suppose the assumptions of Corollary 1.1 hold. Then, for $j = ATE, DTE, CHTE$,

$$\sqrt{nh}\left(\hat{\boldsymbol{v}}^x_j(\cdot,\cdot) - \boldsymbol{v}^x_j(\cdot,\cdot)\right) \Rightarrow v'_{j,\mathbf{S}^x}(\mathbb{G}^x)(\cdot,\cdot),$$

in $\ell_\infty(\mathscr{U} \times \Theta^2) \times \ell_\infty(\mathscr{U} \times \Theta^2)$,[9] where $v'_{j,\mathbf{S}^x}(\mathbb{G}^x)(\cdot,\cdot)$ is a tight, two-dimensional, centered Gaussian process with covariance kernels $\Sigma^x_j(\mathbf{u},\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\psi}^x_j(W,u_1,\boldsymbol{\theta}_1)\boldsymbol{\psi}^x_j(W,u_2,\boldsymbol{\theta}_2)']$, and $\mathbf{u} = (u_1,u_2)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1,\boldsymbol{\theta}'_2)'$. If in addition, for $d \in \{0,1\}$, $0 < \inf_{(\tau,\theta)\in(0,\tau_o)\times\Theta} f_{T_d}(q_{d,\theta}(\tau),\theta) < \sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta} f_{T_d}(q_{d,\theta}(\tau),\theta) < \infty$, we have, in $\ell_\infty(\mathscr{U} \times \Theta^2) \times \ell_\infty(\mathscr{U} \times \Theta^2)$,

$$\sqrt{nh}\left(\hat{\boldsymbol{v}}^x_{QTE}(\cdot,\cdot) - \boldsymbol{v}^x_{QTE}(\cdot,\cdot)\right) \Rightarrow v'_{QTE,\mathbf{S}^x}(\mathbb{G}^x)(\cdot,\cdot).$$

The tight, two-dimensional process $v'_{QTE,\mathbf{S}^x}(\mathbb{G}^x)$ is centered Gaussian with covariance kernel

$$\Sigma^x_{QTE}(\boldsymbol{\tau},\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\psi}^x_{QTE}(W,\tau_1,\boldsymbol{\theta}_1)\boldsymbol{\psi}^x_{QTE}(W,\tau_2,\boldsymbol{\theta}_2)'].$$

(ii) Suppose the assumptions of Corollary 1.2 hold. Then, for $j \in \{ATE, DTE, CHTE\}$,

$$\sqrt{n}\left(\hat{\boldsymbol{v}}_j(\cdot,\cdot) - \boldsymbol{v}_j(\cdot,\cdot)\right) \Rightarrow v'_{j,\mathbf{S}}(\mathbb{G})(\cdot,\cdot),$$

in $\ell_\infty(\mathscr{U} \times \Theta^2) \times \ell_\infty(\mathscr{U} \times \Theta^2)$, where $v'_{j,\mathbf{S}}(\mathbb{G})(\cdot,\cdot)$ is a tight, two-dimensional, centered Gaussian process with covari-

[9] Definition of the set $\mathscr{U}$ depends on the type of treatment effect under consideration. Specifically, $\mathscr{U} = \varnothing$ if $j = ATE$, $\mathscr{U} = (0,\tau_o)$, if $j = QTE$. Otherwise, $\mathscr{U} = \tilde{\mathscr{T}}$.

ance kernels $\Sigma_j(\mathbf{u},\boldsymbol{\theta})=\mathbb{E}[\boldsymbol{\psi}_j(W,u_1,\boldsymbol{\theta}_1)\boldsymbol{\psi}_j(W,u_2,\boldsymbol{\theta}_2)']$, and $\mathbf{u}=(u_1,u_2)$, $\boldsymbol{\theta}=(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2')'$. If in addition, for $d\in\{0,1\}$, $0<\inf_{(\tau,x,\theta)\in(0,\tau_o)\times\mathscr{X}\times\Theta} f_{T_d,x}(q_{d,\theta}^x(\tau),\theta)<\sup_{(\tau,x,\theta)\in(0,\tau_o)\times\mathscr{X}\times\Theta} f_{T_d,x}(q_{d,\theta}^x(\tau),\theta)<\infty$, we have, in $\ell_\infty(\mathscr{U}\times\Theta^2)\times\ell_\infty(\mathscr{U}\times\Theta^2)$,

$$\sqrt{n}\,(\hat{\boldsymbol{\nu}}_{QTE}(\cdot,\cdot)-\boldsymbol{\nu}_{QTE}(\cdot,\cdot))\Rightarrow \nu'_{QTE,\mathbf{S}}(\mathbb{G})(\cdot,\cdot).$$

The tight, two-dimensional process $\nu'_{QTE,\mathbf{S}}(\mathbb{G})$ is centered Gaussian with covariance kernel

$$\Sigma_{QTE}(\boldsymbol{\tau},\boldsymbol{\theta})=\mathbb{E}[\boldsymbol{\psi}_{QTE}(W,\tau_1,\boldsymbol{\theta}_1)\boldsymbol{\psi}_{QTE}(W,\tau_2,\boldsymbol{\theta}_2)'].$$

Theorem 1.4 forms the basis for pointwise as well as uniform inference on the TEBFs. Nonetheless, the result cannot be directly used for such a purpose, since the limit processes contain various unknown quantities. The approximation of these quantities can be avoided by means of a standard nonparametric bootstrap procedure. However, its implementation requires recalculating estimators of $\boldsymbol{\gamma}$, the BGFs, and the TEBFs in each bootstrap iteration. Since optimization of (1.4.1) is computationally intensive, we adopt an alternative multiplier bootstrap procedure that entails approximating the influence functions, but dispense with the need for reoptimizations.

1.5 Multiplier Bootstrap

In this section, we propose simulation methods, based on the multiplier bootstrap, for approximating the limiting processes introduced in the previous section. We show the bootstrapped processes converge uniformly to the limiting Gaussian processes defined in Corollaries 1.1, 1.2, and Theorem 1.4. Given these theoretical results, we then provide practical algorithms for conducting pointwise and uniform inference on the treatment effects.

Let $\{\xi_{i,b}\}_{i=1}^n$ be a sequence of random variables with zero mean and unit variance. We call them the *multiplier weights*. These weights are drawn independently of the main sample $\{W_i\}_{i=1}^n$. Given the weights, we define the following two multiplier processes,

$$\mathbb{G}_{n,\xi}^x(t,\boldsymbol{\theta})\equiv n^{-1/2}h^{1/2}\sum_{i=1}^n \boldsymbol{\varphi}^{x*}(W_i,\xi_i,t,\boldsymbol{\theta}),\ \text{ and }\ \mathbb{G}_{n,\xi}(t,\boldsymbol{\theta})\equiv n^{-1/2}\sum_{i=1}^n \boldsymbol{\varphi}^{*}(W_i,\xi_i,t,\boldsymbol{\theta}),$$

where $\boldsymbol{\varphi}^{x*}\equiv(\eta_{s,1}^*,\eta_{s,0}^*)'$, and $\eta_{s,d}^*(W,\xi,t,\theta)\equiv\xi\cdot\eta_{s,d}(W,t,\theta)$. We also have $\boldsymbol{\varphi}^*\equiv(\varphi_1^*,\varphi_0^*)'$, and $\varphi_d^*(W,\xi,t,\theta)\equiv\xi\left\{\varphi_{d,1}(W,t,\theta)+\varphi_{d,2}(W,t,\theta)\right\}$. Since the influence functions contain unknown quantities, we need to replace them with their estimates in practice. We propose reusing $\hat{G}_d(y,x\hat{\gamma}_d)$ and $\hat{G}_{d,1}(y,x\hat{\gamma}_d)$ in the construction of the influence function estimators:

$$
\begin{aligned}
&\hat{\mathscr{E}}_{d,\hat{\gamma}_d}(y,x) = \mathbb{1}\{D=d\}\left(\mathbb{1}\{Y \le y\} - \hat{G}_d(y,x\hat{\gamma}_d)\right),\\
&\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}(y,x) = \mathbb{1}\{D=d\}\left(R\mathbb{1}\{Y \le y\} - \hat{G}_{d,1}(y,x\hat{\gamma}_d)\right),\\
&\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d}, \hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,x,\theta) = \frac{1}{\phi'_{\theta}\left(\hat{s}_{T,d}(t,x\hat{\gamma}_d,\theta)\right)}\Big\{\phi'_{\theta}\left(\hat{s}_d(t,x\hat{\gamma}_d)\right)\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}(t,x)\\
&\qquad + \frac{1}{n\hat{f}(x\hat{\gamma}_d,d)}\sum_{i=1}^{n} I_{d,t,i}\phi''_{\theta}\left(\hat{s}_d(Y_i,x\hat{\gamma}_d)\right)\left(\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}(Y_i,x) - R_i\hat{\mathscr{E}}_{d,\hat{\gamma}_d}(Y_i,x)\right)K_h(x\hat{\gamma}_d, X_i\hat{\gamma}_d)\Big\}.
\end{aligned}
$$

We define the estimated multiplier processes by substituting the above estimators into the multiplier processes. Specifically,

$$
\hat{\mathbb{G}}^{x}_{n,\xi}(t,\boldsymbol{\theta}) = n^{-1/2}h^{1/2}\sum_{i=1}^{n}\hat{\boldsymbol{\varphi}}^{x*}(W_i,\xi_i,t,\boldsymbol{\theta}), \text{ and } \hat{\mathbb{G}}_{n,\xi}(t,\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\hat{\boldsymbol{\varphi}}^{*}(W_i,\xi_i,t,\boldsymbol{\theta}),
$$

where $\hat{\boldsymbol{\varphi}}^{x*} = (\hat{\varphi}^{x*}_1, \hat{\varphi}^{x*}_0)'$, $\hat{\boldsymbol{\varphi}}^{*} = (\hat{\varphi}^{*}_1, \hat{\varphi}^{*}_0)'$, $\hat{\varphi}^{x*}_d(W,\xi,t,\theta) = \xi \cdot \hat{\eta}_{s,d}(W,x,t,\theta)$, $\hat{\varphi}^{*}_d(W,\xi,t,\theta) = \xi \cdot \{\hat{\varphi}_{d,1}(W,t,\theta) + \hat{\varphi}_{d,2}(W,t,\theta)\}$, and

$$
\hat{\eta}_{s,d}(W,\xi,x,t,\theta) = \frac{K_h(x\hat{\gamma}_d, X\hat{\gamma}_d)}{\hat{f}(x\hat{\gamma}_d,d)}\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d}, \hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,x,\theta), \tag{1.5.1}
$$

$$
\hat{\varphi}_{d,1}(W,t,\theta) = \frac{\hat{f}(X\hat{\gamma}_d)}{\hat{f}(X\hat{\gamma}_d,d)}\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d}, \hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,X,\theta), \tag{1.5.2}
$$

$$
\hat{\varphi}_{d,2}(W,t,\theta) = \hat{s}_{T,d}(t,X\hat{\gamma}_d,\theta) - \mathbb{E}_n[\hat{s}_{T,d}(t,X\hat{\gamma}_d,\theta)]. \tag{1.5.3}
$$

Assumption 1.11 (Multiplier weights) $\{\xi_i\}_{i=1}^{n}$ is a sequence of *i.i.d.* random variables, defined on a probability space independent of $\{W_i\}_{i=1}^{n}$, satisfying $E[\xi_1] = 0$ and $\mathbb{E}[\xi_1^2] = 1$.

There are several different choices for ξ that are commonly encountered in the literature. For instance, when $\xi = \mathscr{N}$, where $\mathscr{N}$ is a standard normal random variable, it is referred to as the Gaussian multiplier method, as seen in Giné and Zinn (1984). When $\xi = \mathscr{N}_1/\sqrt{2} + (\mathscr{N}_2^2 - 1)/2$, where $\mathscr{N}_1$ and $\mathscr{N}_1$ are mutually independent standard normal random variables, it corresponds to the wild bootstrap method, as seen in Mammen (1993).

In the next theorem, we show that the estimated multiplier processes $\hat{\mathbb{G}}^{x}_{n,\xi}$ and $\hat{\mathbb{G}}_{n,\xi}$ approximate $\mathbb{G}^{x}$ and $\mathbb{G}$, respectively. The approximation, formally termed as *conditional weak convergence in probability*, where the condition is on the main sample, is in the sense of Section 2.2.3 in Kosorok (2008).

Theorem 1.5 Under the assumptions of Theorem 1.3, Assumptions 1.10, and 1.11, we have that (i) $\hat{\mathbb{G}}^{x}_{n,\xi} \underset{\xi}{\overset{p}{\rightsquigarrow}} \mathbb{G}^{x}$, and (ii) $\hat{\mathbb{G}}_{n,\xi} \underset{\xi}{\overset{p}{\rightsquigarrow}} \mathbb{G}$.

Theorem 1.5, combined with the functional delta method for the bootstrap (see e.g. Theorem 3.9.11 in Van Der Vaart and Wellner (1996)), allows us to establish the validity of plug-in estimators of Hadamard differentiable functionals. Let us first define the estimated multiplier processes for the bound curves, $\hat{\mathbb{G}}_{\xi,j}$ and $\hat{\mathbb{G}}^x_{\xi,j}$, by

$$\hat{\mathbb{G}}^x_{\xi,j}(u,\boldsymbol{\theta}) = n^{-1/2}h^{1/2}\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}^{x*}_j(W_i,\xi_i,u,\boldsymbol{\theta}), \quad \text{and} \quad \hat{\mathbb{G}}_{\xi,j}(u,\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}^{*}_j(W_i,\xi_i,u,\boldsymbol{\theta}), \tag{1.5.4}$$

for $j \in \{ATE, DTE, QTE, CHTE\}$, $\boldsymbol{\theta} = (\theta_1,\theta_2)' \in \Theta^2$, with $\theta_1 \leq \theta_2$. In the preceding definition, we use the following estimators of the influence functions: $\hat{\boldsymbol{\psi}}^*_j(W,\xi,u,\boldsymbol{\theta}) \equiv \xi \cdot (\hat{\psi}_{1,j}(W,u,\theta_2) - \hat{\psi}_{0,j}(W,u,\theta_1), \hat{\psi}_{1,j}(W,u,\theta_1) - \hat{\psi}_{0,j}(W,u,\theta_2))'$ and $\hat{\boldsymbol{\psi}}^{x*}_j(W,\xi,u,\boldsymbol{\theta}) \equiv \xi \cdot (\hat{\psi}^x_{1,j}(W,u,\theta_2) - \hat{\psi}^x_{0,j}(W,u,\theta_1), \hat{\psi}^x_{1,j}(W,u,\theta_1) - \hat{\psi}^x_{0,j}(W,u,\theta_2))'$, $j \in \{ATE, QTE\}$; moreover, $\hat{\boldsymbol{\psi}}^*_j(W,\xi,u,\boldsymbol{\theta}) \equiv \xi \cdot (\hat{\psi}_{1,j}(W,u,\theta_1) - \hat{\psi}_{0,j}(W,u,\theta_2), \hat{\psi}_{1,j}(W,u,\theta_2) - \hat{\psi}_{0,j}(W,u,\theta_1))'$ $\hat{\boldsymbol{\psi}}^{x*}_j(W,\xi,u,\boldsymbol{\theta}) \equiv \xi \cdot (\hat{\psi}^x_{1,j}(W,u,\theta_1) - \hat{\psi}^x_{0,j}(W,u,\theta_2), \hat{\psi}^x_{1,j}(W,u,\theta_2) - \hat{\psi}^x_{0,j}(W,u,\theta_1))'$, for $j \in \{DTE, CHTE\}$, where

$$\hat{\psi}_{d,ATE}(W,t,\theta) = \sum_{i=1}^{n-1}\mathbb{1}\left\{Y_{(i+1):n} \leq t\right\}\left(Y_{(i+1):n} - Y_{i:n}\right)\hat{\varphi}_d(W,Y_{i:n},\theta), \tag{1.5.5}$$

$$\hat{\psi}^x_{d,ATE}(W,t,\theta) = \sum_{i=1}^{n-1}\mathbb{1}\left\{Y_{(i+1):n} \leq t\right\}\left(Y_{(i+1):n} - Y_{i:n}\right)\hat{\eta}_{s,d}(W,x,Y_{i:n},\theta), \tag{1.5.6}$$

$$\hat{\psi}_{d,DTE}(W,t,\theta) = -\hat{\varphi}_d(W,t,\theta), \qquad \hat{\psi}^x_{d,DTE}(W,t,\theta) = -\hat{\eta}_{s,d}(W,x,t,\theta), \tag{1.5.7}$$

$$\hat{\psi}_{d,QTE}(W,\tau,\theta) = \frac{\hat{\varphi}_d(W,\hat{q}_{d,\theta}(\tau),\theta)}{\hat{f}_{T_d}(\hat{q}_{d,\theta}(\tau),\theta)}, \qquad \hat{\psi}^x_{d,QTE}(W,\tau,\theta) = \frac{\hat{\eta}_{s,d}(W,x,\hat{q}^x_{d,\theta}(\tau),\theta)}{\hat{f}_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)}, \tag{1.5.8}$$

$$\hat{\psi}_{d,CHTE}(W,t,\theta) = -\frac{\hat{\varphi}_d(W,t,\theta)}{\hat{s}_{T_d}(t,\theta)}, \qquad \hat{\psi}^x_{d,CHTE}(W,t,\theta) = -\frac{\hat{\eta}_{s,d}(W,x,t,\theta)}{\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta)}, \tag{1.5.9}$$

for $d \in \{0,1\}$. In (1.5.8), $\hat{f}_{T_d}(t,\theta)$ and $\hat{f}_{T_d,x}(t,\theta)$ are any first-stage estimators of $f_{T_d}(t,\theta) \equiv -\partial s_{T_d}(y,\theta)/\partial y|_{y=t}$ and $f_{T_d,x}(t,\theta) \equiv -\partial s_{T_d}(y,x\gamma_d,\theta)/\partial y|_{y=t}$, respectively, that are uniformly convergent as required in the assumption below.

Assumption 1.12 (First stage density estimator) There exist first stage estimators $\hat{f}_{T_d}(t,\theta)$ and $\hat{f}_{T_d,x}(t,\theta)$ that are consistent for $f_{T_d}(t,\theta)$ and $f_{T_d,x}(t,\theta)$, respectively, uniformly over $\tilde{\mathscr{T}} \times \mathscr{X} \times \Theta$, for $d \in \{0,1\}$.

In Section 1.9.3.4, we describe an estimator of the conditional density and show that it fulfills the preceding assumption.

Corollary 1.3 Suppose the assumptions of Theorem 1.4, Assumptions 1.11, and 1.12 hold, we get (i) $\hat{\mathbb{G}}^x_{\xi,j} \underset{\xi}{\overset{p}{\rightarrow}} \nu'_{j,\mathbf{S}^x}(\mathbb{G}^x)$, and (ii) $\hat{\mathbb{G}}_{\xi,j} \underset{\xi}{\overset{p}{\rightarrow}} \nu'_{j,\mathbf{S}}(\mathbb{G})$, for $j \in \{ATE, DTE, QTE, CHTE\}$.

1.5.1 Bootstrap Confidence Bands

The functional central limit theorems for multiplier bootstrap established in the previous section can be used to conduct point-wise and uniform inference for the TEBF estimators. We provide an algorithm for constructing uniform confidence bands of the overall TEBF estimators in what follows. An analogous procedure that produces uniform confidence bands of the conditional TEBF estimators is given in Section 1.9.1.3.2. Point-wise confidence intervals are by-products of these two algorithms.

Let $\hat{\mathbb{G}}_{lb,\xi,j}$ and $\hat{\mathbb{G}}_{ub,\xi,j}$ denote the first and second component of $\hat{\mathbb{G}}_{\xi,j}$, respectively.

Algorithm 1.5.1 (Uniform confidence sets of overall TEBFs)

1. Select a finite grid set $\mathscr{U}_m \equiv \{u_1, u_2, ..., u_m\}$ from $\mathscr{U}$, where the index u depends on the type of treatment effect under consideration. Pick a set $\boldsymbol{\Theta}_l \equiv \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_l\}$ with $\boldsymbol{\theta}_s = (\theta_{1,s}, \theta_{2,s}) \in \Theta^2$, and $\theta_{1,s} \leq \theta_{2,s}$, for all $s = 1, ..., l$.

   In Steps 2-5, the calculations will be performed for $d, r \in \{0,1\}$, $t \in \tilde{\mathscr{T}}$, $\tau \in (0, \tau_o)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_l$, and $u \in \mathscr{U}_m$.

2. Estimate $\hat{\gamma}_d$, $\hat{G}_{d,1}(t, x\hat{\gamma}_d)$, $\hat{G}_d(t, x\hat{\gamma}_d)$, and $\hat{s}_{T_d}(t,\theta)$. If $j = QTE$, compute $\hat{q}_{d,\theta}(\tau)$ and $\hat{f}_{T_d}(t,\theta)$.

3. Calculate $\hat{\nu}_j(u,\boldsymbol{\theta})$, $\hat{\varphi}_{d,1}(W,t,\theta)$, $\hat{\varphi}_{d,2}(W,t,\theta)$, and $\hat{\psi}_j(W,\theta)$.

4. Sample $\{\xi_i^b\}_{i=1}^n$ from a distribution with zero mean and unit variance, independently from data. Calculate $\hat{\boldsymbol{\psi}}_j^*$, and $\mathbb{G}_{\xi^b,j}(u,\theta)$.

   Repeat Step 4 for $b = 1, ..., B$, where B is some large integer.

5. For $\ell = lb, ub$, compute the $(1-\alpha)$-th quantile $\hat{c}^B_{n,\ell,j}(\alpha, \mathscr{U}_m, \boldsymbol{\Theta}_l)$ of $\left\{\max_{1\leq i\leq m, 1\leq s\leq l}\left\|\mathbb{G}_{\ell,\xi^b,j}(u_i,\boldsymbol{\theta}_s)\right\|\right\}_{b=1}^B$, and construct the uniform confidence band

$$C^B_{n,\ell,j}(1-\alpha, \mathscr{U}_m, \boldsymbol{\Theta}_l) \equiv \left\{\hat{\nu}_{\ell,j}(u,\boldsymbol{\theta}) \pm n^{-1/2}\hat{c}^B_{n,\ell,j}(\alpha,\mathscr{U}_m,\boldsymbol{\Theta}_l) : u \in \mathscr{U}_m, \boldsymbol{\theta} \in \boldsymbol{\Theta}_l\right\}.$$

The null hypothesis, which posits that a given TEBF is identically 0 over the index set $\mathscr{U}_m$, can be tested directly using the simulated bootstrap critical values. More generally, tests of FOSD relations, such as $\{\nu_{j,\ell}(u,\boldsymbol{\theta}) \leq 0 : u \in \mathscr{U}_m, \boldsymbol{\theta} \in \boldsymbol{\Theta}_l\}$, and tests of homogeneity, such as $\{\nu_{j,\ell}(u,\boldsymbol{\theta}) = \int_{\mathscr{U}_m} \nu_{j,\ell}(\tilde{u},\boldsymbol{\theta})d\tilde{u} : u \in \mathscr{U}_m, \boldsymbol{\theta} \in \boldsymbol{\Theta}_l\}$ can also be easily constructed using the simulated bootstrap process $\left\{\mathbb{G}_{\ell,\xi^b,j}\right\}_{b=1}^B$, which is a byproduct of Algorithm 1.5.1.

Note that $C^B_{n,\ell,j}(1-\alpha,\mathscr{U}_m,\boldsymbol{\Theta}_l)$ are uniform across both u and θ. Pointwise confidence sets are immediately available from the two aforementioned procedures, by setting $\mathscr{U}_m = \{u^*\}$, $\boldsymbol{\Theta}_l = \{\boldsymbol{\theta}^*\}$, for some u^* and $\boldsymbol{\theta}^*$ specified by the researcher.

We denote the confidence set generated by Algorithm 1.9.1 as $C^{x,B}_{n,\ell,j}(1-\alpha,\mathscr{U}_m,\boldsymbol{\Theta}_l)$. The next theorem confirms that the uniform bootstrap confidence bands for both conditional and overall TEBFs are asymptotically accurate.

Theorem 1.6 Suppose the assumptions of Corollary 1.3 hold, we have

$$\lim_{n\to\infty}\inf_{(u,\boldsymbol{\theta})\in\mathscr{U}_m\times\boldsymbol{\Theta}_l}\mathbb{P}\left(\nu^x_{\ell,j}(u,\boldsymbol{\theta})\in C^{x,B}_{n,\ell,j}(1-\alpha,\mathscr{U}_m,\boldsymbol{\Theta}_l)\right)=1-\alpha,$$

$$\lim_{n\to\infty}\inf_{(u,\boldsymbol{\theta})\in\mathscr{U}_m\times\boldsymbol{\Theta}_l}\mathbb{P}\left(\nu_{\ell,j}(u,\boldsymbol{\theta})\in C^{B}_{n,\ell,j}(1-\alpha,\mathscr{U}_m,\boldsymbol{\Theta}_l)\right)=1-\alpha,$$

for $x\in\mathscr{X}$, $\ell\in\{lb,ub\}$, and $j\in\{ATE,DTE,QTE,CHTE\}$.

1.6 Monte Carlo Study

Results from the previous section imply that the estimators and uniform confidence bands for the conditional and overall TEBFs will exhibit desirable properties when sample size is sufficiently large. But what about their small-sample performance? To address this question, we conducted a small scale Monte Carlo experiment. The DGP for the simulations consists of the following four aspects:

1. The conditional survival functions: For $d\in\{0,1\}$, both T_d and C_d follow the conditional Exponential distribution. Specifically, $S_{\ell_d|X}(t|x)=\exp\left(-\lambda_{\ell,d}(x\gamma_d)t\right)$, where the hazard rate parameter $\lambda_{\ell,d}(x\gamma_d)>0$, for $x\in\mathscr{X}$, $\ell\in\{T,C\}$, and $d\in\{0,1\}$.

2. The copula function: The true conditional survival copula is assumed to belong to the Archimedean family with the Gumbel generator function: $\phi_\theta(u)=\left(\log u^{-1}\right)^{\theta}$ and

$$\mathscr{C}_x(u,v)=\exp\left(-\left[(\log u^{-1})^{\theta^*_d(x)}+(\log v^{-1})^{\theta^*_d(x)}\right]^{1/\theta^*_d(x)}\right),$$

where $\theta^*_d(x)\subset[1,\infty)$, for all $x\in\mathscr{X}$ and $d\in\{0,1\}$.

3. The covariates: $X=(X_1,X_2,X_3)$, where X_1 and X_2 are drawn from truncated normal distribution with mean and standard deviation equal to 0.5 and 1, respectively. The two variables are restricted to lie in $[0.01,1]$. The remaining one X_3, is a binary variable, following Bernoulli(0.5). The three variables are mutually independent.

4. The treatment assignment mechanism: the treatment status D is determined by $D=\mathbb{1}\{p(X)>U\}$, where U follows Uniform[0, 1], and

$$p(X)=\frac{\exp(-0.1X_1+0.1X_2)}{1+\exp(-0.1X_1+0.1X_2)},$$

is the true propensity score function.

The BGF admits an analytical form when $\lambda_{T,d}(\cdot) = \lambda_{C,d}(\cdot) \equiv \lambda_d(\cdot)$. Such simplification is handy when checking the coverage of our bootstrap confidence sets. By symmetry, we have that $S_{Y_d|X}(y|x) = \exp\left(-2^{1/\theta_d^*(x)}\lambda_d(x\gamma_d)y\right)$, and $S_{Y_d,R_d|X}(y,1|x) = 2^{-1}\exp\left(-2^{1/\theta_d^*(x)}\lambda_d(x\gamma_d)y\right)$, implying a population censoring rate of 50%. Now, from (1.4.5) and by direct calculations,

$$s_{T_d}(t, x\gamma_d, \theta) = \phi_\theta^{-1}\left(2^{-1}\phi_\theta\left(\exp\left(-2^{1/\theta_d^*(x)}\lambda_d(x\gamma_d)t\right)\right)\right), \text{ for } d \in \{0,1\}.$$

This formula simplifies further when the true copula is Gumbel. In this case, $s_{T_d}(t, x\gamma_d, \theta) = \exp(-2^{1/\theta_d^*(x)-1/\theta} \cdot \lambda_d(x\gamma_d)t)$, equivalent to an exponential distribution with a rate parameter equal to $\beta_d(x\gamma_d, \theta) \equiv 2^{1/\theta_d^*(x)-1/\theta}\lambda_d(x\gamma_d)$. As a direct consequence, $v_{ATE,lb}^x(\boldsymbol{\theta}) = \beta_1(x\gamma_1, \theta_2)^{-1} - \beta_0(x\gamma_0, \theta_1)^{-1}$, and the overall average effect is also immediately available via taking the expectation with respect to X. In the following, we set $\gamma_1 = \gamma_0 = (0.5, 0.5)'$, $\lambda_1(v) = \sqrt{v} + v/2$, and $\lambda_0(v) = \sqrt{v}$. As a result, the true DTE is heterogeneous and uniformly negative across the index set.

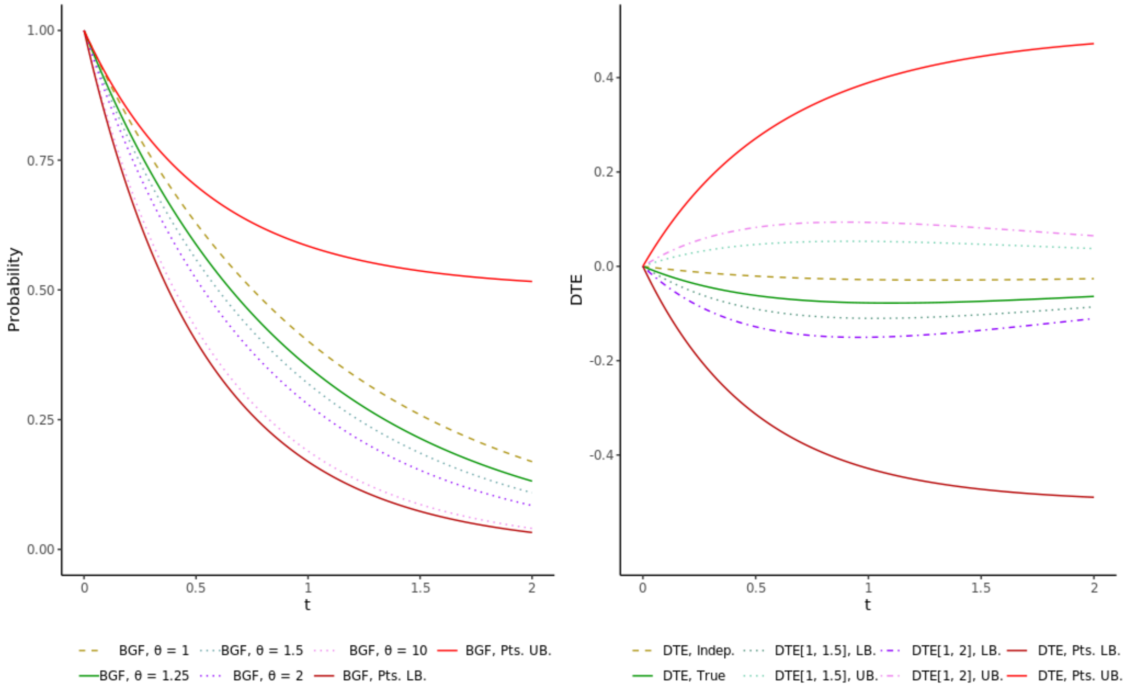
To generate variables from the Gumbel copula, we follow the algorithm provided in Section 2.9 in Nelsen (2007), for which purpose, we assume the true copula parameters are $\theta_0^*(\cdot) = 1$ and $\theta_1^*(\cdot) = 1.25$. That is, censoring is independent for the treated group, whereas T_0 and C_0 are correlated with Kendall's τ equal to 0.2.

In Figure 1.1, we plot unconditional BGFs and DTEs across various levels of the sensitivity parameter, alongside the corresponding Peterson's bounds. We observe that the worse case bounds (the upper bound in particular) are highly non-informative and the BGFs provide significant improvement over the worst-case bounds under the assumed censoring mechanism. One may question whether the gap between the two can be completely bridged by varying theta. The answer to this question is contingent on the copula under consideration and specifically whether it admits Hoeffding-Frechet bounds as limiting cases. For instance, the Gumbel copula is unable to bridge the gap entirely, due to its inherent incapability to model negative correlation. Additionally, the figure illustrates the stochastic dominance relations, as influenced by the concordance ordering within the copula family. Consequently, we can observe a correlation between the size of the identified set and the range of theta values chosen by the researcher.

To assess the performance of TEBF estimators over an index set $\mathscr{U}_m$, we adopt *average* and *median integrated bias*, *integrated root mean square error* (IRMSE), and the coverage rate as the criterion of evaluation.[10] Regarding the index set $\mathscr{U}_m$, we use an equidistant grid between 0.1 and 1.5 with the interval size of 0.05 for the ATE, DTE and CHTE. For the QTE, an equidistant grid between 0.25 and 0.75 with a step size of 0.05 is adopted. We let both

[10] Consider a Monte Carlo experiment with S replications, the average integrated bias is defined by $S^{-1}\sum_{s=1}^{S}\int_{u\in\mathscr{U}_m}\left|\hat{f}_s(u) - f(u)\right|du$, median integrated bias denotes the 50-th percentile of $\left\{\int_{u\in\mathscr{U}_m}\left|\hat{f}_s(u) - f(u)\right|du\right\}_{s=1}^{S}$, and the IRMSE, by $\left(S^{-1}\sum_{s=1}^{S}\int_{u\in\mathscr{U}_m}\left|\hat{f}_s(u) - f(u)\right|^2 du\right)^{1/2}$, where $f(\cdot)$ is any one of the $v_{j,\ell}(\cdot)$, for $\ell = lb, ub$, and $j \in \{ATE, DTE, QTE, CHTE\}$.

Figure 1.1: BGFs and DTEs with multiple levels of θ



Notes: The left panel depicts the unconditional BGFs for the control group and the right panel illustrates the overall DTEs. In each plot, the dashed curve depicts the function when the independent censoring mechanism is assumed (equivalently, Gumbel copula with $\theta = 1$). The green solid curves represent the true functions (Gumbel copula with $\theta = 1.25$). The red solid curves depict the Peterson's worst-case bounds.

$L(\cdot)$ and $K(\cdot)$ be the Epanechnikov kernel: $L(u) = K(u) = 0.75(1-u^2)\mathbb{1}\{|u| \leq 1\}$. For treatment group $d \in \{0,1\}$, the bandwidth b is chosen as the value from the set $\left\{2^{-0.5k}(n/2)^{-0.26}\right\}_{k=-1}^{6}$, that minimizes the estimated criterion $\hat{\mathscr{J}}_d(\hat{\gamma}_d, \rho)$, where we let $\rho(v) = \exp(-\|v\|^2/2)$. We then set the bandwidth h equal to b. To assess the impact of first-step estimation, we provide a set of "oracle" results where the single index parameters take their true values along with "feasible" results where the parameters are estimated according to the procedure from Section 1.4.1.

Table 1.1 reports simulation results based on 1,000 Monte Carlo replications of samples with size $n = 1,000$. For each type of treatment effect, we show results for two different range of θ: a narrower one with $\boldsymbol{\theta} = (1, 1.5)$, and a wider one with $\boldsymbol{\theta} = (1,2)$. Given the one-to-one mapping between θ and Kendall's τ, the two $\boldsymbol{\theta}$ choices correspond to Kendall's τ lying between $[0, 1/3]$ and $[0, 1/2]$, respectively. Results in Table 1.1 suggest that our TEBF estimators exhibit minimal bias, and their confidence intervals generally achieve close-to-nominal-level coverage, irrespective of the choice of copula parameters, the type of treatment effects and whether the effect is conditional. The (C)CHTE perform relatively worse than the other three types, in terms of integrated bias and IRMSE. This is to be expected as the log transformations tend to induce higher bias.

When comparing oracle and feasible results, we find that the oracle results generally exhibit smaller bias and

IRMSE, and they have better coverage properties. The difference is more prominent when conditional treatment effects are considered. This is partially due to the fact that the component in the expansion of the conditional TEBFs, which is associated with the single-index estimation, is of a lower order than the component appearing in the overall TEBFs, even though both are negligible in the first order. To improve the performance of our bootstrap procedure for conditional TEBFs, one may consider adding the influence functions associated with first stage estimation when constructing the bootstrap processes. This is left for future research.

Overall, the results from finite-sample studies align with the theoretical predictions discussed in Sections 1.4 and 1.5.

1.7 Empirical Illustration

In this section, we revisit Bernasconi et al. (2022) on the effect of acute lymphoblastic leukaemia treatment where survival time is subject to dependent censoring caused by the abandonment of treatment. ALL is a major cause of cancer diagnoses among people under 18 years old, accounting for nearly 25% of all cancer diagnoses (Howlader et al., 2016). Wide disparities in cure rates has been documented between high-income (approximately 80%) and mid-and-low-income (approximately 35%) countries (Gatta et al., 2005; Howard and Wilimas, 2005). Abandonment of treatment is seen as a major factor for such disparities (Mostert et al., 2011). Decision to withdraw treatment can be affected by various factors including distance to the treatment facility, family economic status, and personal beliefs, many of which also have an impact on the quality of treatment. As such, the independent censoring assumption is not appropriate in this context.

The data comes from two subsequent clinical studies conducted in Honduras between 2000 and 2015. During the period from 2000 to 2007, a protocol called GHS-2000 were adopted to treat ALL patients. In the second period (2008–2015), the treatment follows a new protocol denominated AHOPCA ALL-2008.[11] We view GHS-2000 as the control group ($d = 0$) and AHOPCA ALL-2008 as the treated group ($d = 1$). Treatment effects in this context translate to the comparative effectiveness of the two protocols. The outcome of interest T is formally defined as the time since treatment to the first event among relapse, resistance to treatment, secondary malignant neoplasm, and death. The outcome is subject to both administrative censoring, which is independent of the EFS time, and endogenous censoring in the form of abandonment of treatment. Following Bernasconi et al. (2022), we combine the two types of censoring into a composite variable C, and we use potential *event-free-survival* (EFS) to denote S_{T_d} under protocol $d \in \{0, 1\}$.

The baseline covariates of the study include biological characteristics such as gender, age, white blood cell count, central nervous system involvement (CNS), cancer histology, and socio-economic factors: family unity, living con-

[11] Details of these two protocols can be found in Marjerrison et al. (2013) and Navarrete et al. (2014)

Table 1.1: Monte Carlo results for the conditional and overall TEBFs

(a)	Lower Conditional TEBF					Upper Conditional TEBF			
Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
CATE	0.040	0.032	0.051	0.929	CATE	0.040	0.032	0.051	0.934
CDTE	0.072	0.062	0.081	0.933	CDTE	0.073	0.063	0.082	0.928
CQTE	0.066	0.057	0.133	0.924	CQTE	0.067	0.057	0.134	0.936
CCHTE	0.186	0.151	0.259	0.937	CCHTE	0.186	0.151	0.259	0.937
Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
CATE	0.026	0.021	0.033	0.938	CATE	0.025	0.020	0.032	0.942
CDTE	0.046	0.040	0.051	0.947	CDTE	0.047	0.039	0.051	0.927
CQTE	0.043	0.036	0.086	0.932	CQTE	0.045	0.039	0.089	0.938
CCHTE	0.114	0.096	0.138	0.938	CCHTE	0.114	0.096	0.138	0.938

(b)	Lower Conditional TEBF					Upper Conditional TEBF			
Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
CATE	0.039	0.032	0.051	0.932	CATE	0.039	0.031	0.050	0.936
CDTE	0.070	0.061	0.079	0.927	CDTE	0.071	0.062	0.080	0.923
CQTE	0.062	0.053	0.125	0.925	CQTE	0.063	0.054	0.127	0.930
CCHTE	0.192	0.157	0.267	0.939	CCHTE	0.190	0.156	0.259	0.943
Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
CATE	0.025	0.020	0.032	0.936	CATE	0.025	0.020	0.032	0.938
CDTE	0.046	0.040	0.050	0.948	CDTE	0.046	0.039	0.050	0.929
CQTE	0.040	0.034	0.081	0.929	CQTE	0.044	0.038	0.085	0.941
CCHTE	0.118	0.100	0.144	0.939	CCHTE	0.119	0.102	0.144	0.940

(c)	Lower Overall TEBF					Upper Overall TEBF			
Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
ATE	0.023	0.018	0.028	0.958	ATE	0.023	0.019	0.028	0.956
DTE	0.041	0.036	0.044	0.965	DTE	0.041	0.036	0.044	0.963
QTE	0.037	0.033	0.074	0.961	QTE	0.038	0.032	0.074	0.965
CHTE	0.103	0.091	0.122	0.966	CHTE	0.102	0.088	0.120	0.967
Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
ATE	0.021	0.018	0.026	0.954	ATE	0.021	0.018	0.026	0.963
DTE	0.039	0.034	0.042	0.954	DTE	0.039	0.035	0.042	0.950
QTE	0.036	0.031	0.071	0.948	QTE	0.038	0.034	0.074	0.949
CHTE	0.097	0.085	0.115	0.951	CHTE	0.096	0.085	0.114	0.957

(d)	Lower Overall TEBF					Upper Overall TEBF			
Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Feasible	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
ATE	0.022	0.018	0.028	0.963	ATE	0.022	0.019	0.027	0.957
DTE	0.040	0.035	0.042	0.965	DTE	0.040	0.035	0.043	0.966
QTE	0.035	0.031	0.069	0.968	QTE	0.036	0.032	0.071	0.965
CHTE	0.107	0.096	0.127	0.965	CHTE	0.104	0.091	0.123	0.968
Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate	Oracle	Avg. Intg. Bias	Med. Intg. Bias	IRMSE	Cvg. Rate
ATE	0.021	0.017	0.026	0.954	ATE	0.021	0.017	0.025	0.966
DTE	0.038	0.034	0.041	0.955	DTE	0.038	0.034	0.041	0.954
QTE	0.034	0.030	0.068	0.957	QTE	0.036	0.032	0.071	0.958
CHTE	0.100	0.088	0.119	0.952	CHTE	0.099	0.087	0.118	0.957

Notes: Simulations are based on 1,000 Monte Carlo experiments with samples of size $n = 1,000$. Panels (a) and (c) present results for conditional and overall TEBF with $\boldsymbol{\theta} = (1, 1.5)$. Panels (b) and (d) correspond to $\boldsymbol{\theta} = (1, 2)$. In each panel, "feasible" represents results generated with the single-index parameters estimated following Section 1.4.1, whereas the results in the "oracle" sub-panel correspond to those generated using the true single-index parameters. "Avg. Intg. Bias", "Med. Intg. Bias", "IRMSE", and "Cvg. Rate" stand for the average integrated bias, median integrated bias, integrated root mean squared errors, and 95% empirical coverage probability, respectively. The empirical coverage probability is based on bootstrap confidence sets computed with 1,999 multiplier bootstrap replications.

Table 1.2: Summary statistics

	GHS-2000 (No. Obs. = 514)					AHOPCA ALL-2008 (No. Obs. = 536)				
Statistics	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
Follow-up Duration, Years	4	3.84	0.47	2.73	7.59	2.29	2.05	0.71	1.55	3.47
Abandonments	21%	0.41	0	0	0	15%	0.35	0	0	0
Age	7.67	4.36	3.79	6.79	11.26	7.67	4.94	3.45	6.01	11.82
White Blood Cell Count	4.64	10.01	0.45	1.13	4.24	4.28	8.52	0.53	1.11	4.07
Time to Hospital	4.05	2.85	1.5	4	6	2.87	2.03	1	3	4
Male	61%	0.49	0	1	1	53%	0.5	0	1	1
CNS	6%	0.23	0	0	0	13%	0.34	0	0	0
Linage	92%	0.27	1	1	1	93%	0.26	1	1	1
Living Condition	62%	0.49	0	1	1	46%	0.5	0	0	1
Family Type	45%	0.5	0	0	1	14%	0.35	0	0	0
Phone at Home	36%	0.48	0	0	1	42%	0.49	0	0	1

Note: Summary statistics for two protocols for ALL treatment. The left panel describes the GHS-2000 group (2000 - 2007). The right panel is for the AHOPCA ALL-2008 group (2008-2015). ”MALE”, ”CNS”, ”Lineage”, ”Living Condition”, ”Family Type”, ”Phone at Home” are the dummy variables. These variables stand for whether the subject is male, the involvement of the central nervous system, the type of tumor lineage, whether the patient lives in an urban neighborhood, whether the patient lives in a united family, and if the patient owns a home phone.

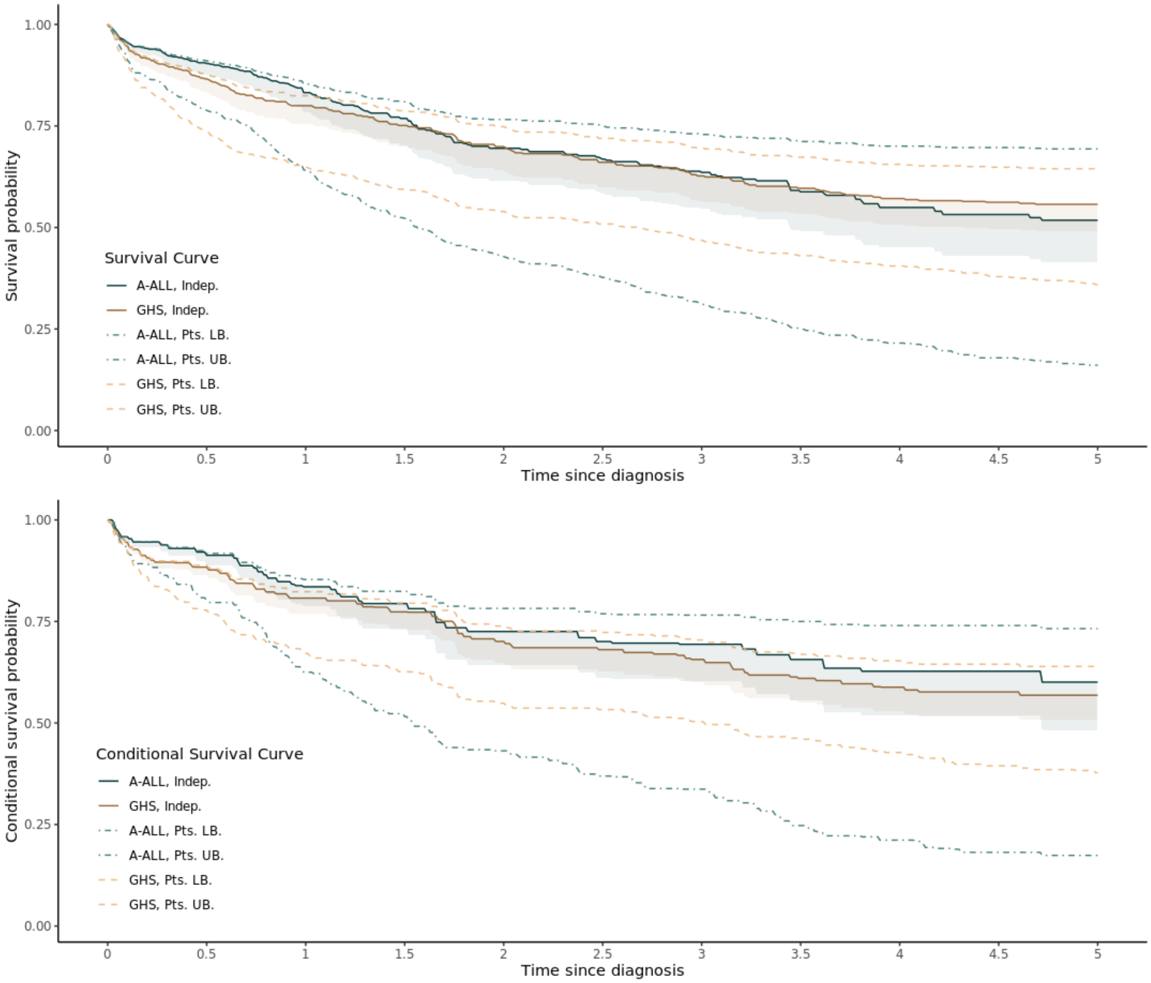
ditions, home phone ownership, and distance to the hospital. Table 1.2 summarizes these characteristics for patients undergoing each of the two protocols. Instead of performing multiple imputations as in Bernasconi et al. (2022), missing cases are removed. Results from the table show that patients from the AHOPCA ALL-2008 group are less likely to withdraw treatment, more likely to live in a rural neighborhood, and tend to live closer to the clinic. To account for these imbalances across the treatment groups, Bernasconi et al. (2022) relies on an *inverse probability of treatment and censoring* weighting strategy, which further depends on the assumption that, conditional on the baseline covariates, the potential EFS and abandonment are mutually independent. Such a restriction, however, is not necessary for our proposed methodology.

The main finding of Bernasconi et al. (2022) is that the AHOPCA ALL-2008 protocol leads to better potential EFS in the first three years and the difference tapers off in the long term (approximately 5 years). Could these results carry over to the case of dependent censoring? To address this question, we consider two scenarios that are characterized by different ranges of the copula parameter θ. Specifically, we assume that the true copula belongs to the Gumbel family and the indexing parameters lie in $(1, 1.5)$ for the first case and in $(1, 2)$ for the second. Both scenarios feature a mild positive correlation pattern between the EFS and withdrawal time, and both encompass independent censoring as a limiting case. When mapped to Kendall's τ, the maximum levels of positive correlation under the two scenarios are $1/3$ and $1/2$, respectively.

In the first step of analysis, we assess the validity of the index sufficiency assumption. In light of Remark 1.1, we can implement the specification test (Algorithm 1.9.2) as presented in Section 1.9.2.2. The null hypothesis is that the single index assumption holds for the joint distribution of (T_d, C_d), $d \in \{0,1\}$. The bootstrap test cannot reject the null for either treatment group at the 10% level, indicating that our methodology can be applied to this context.

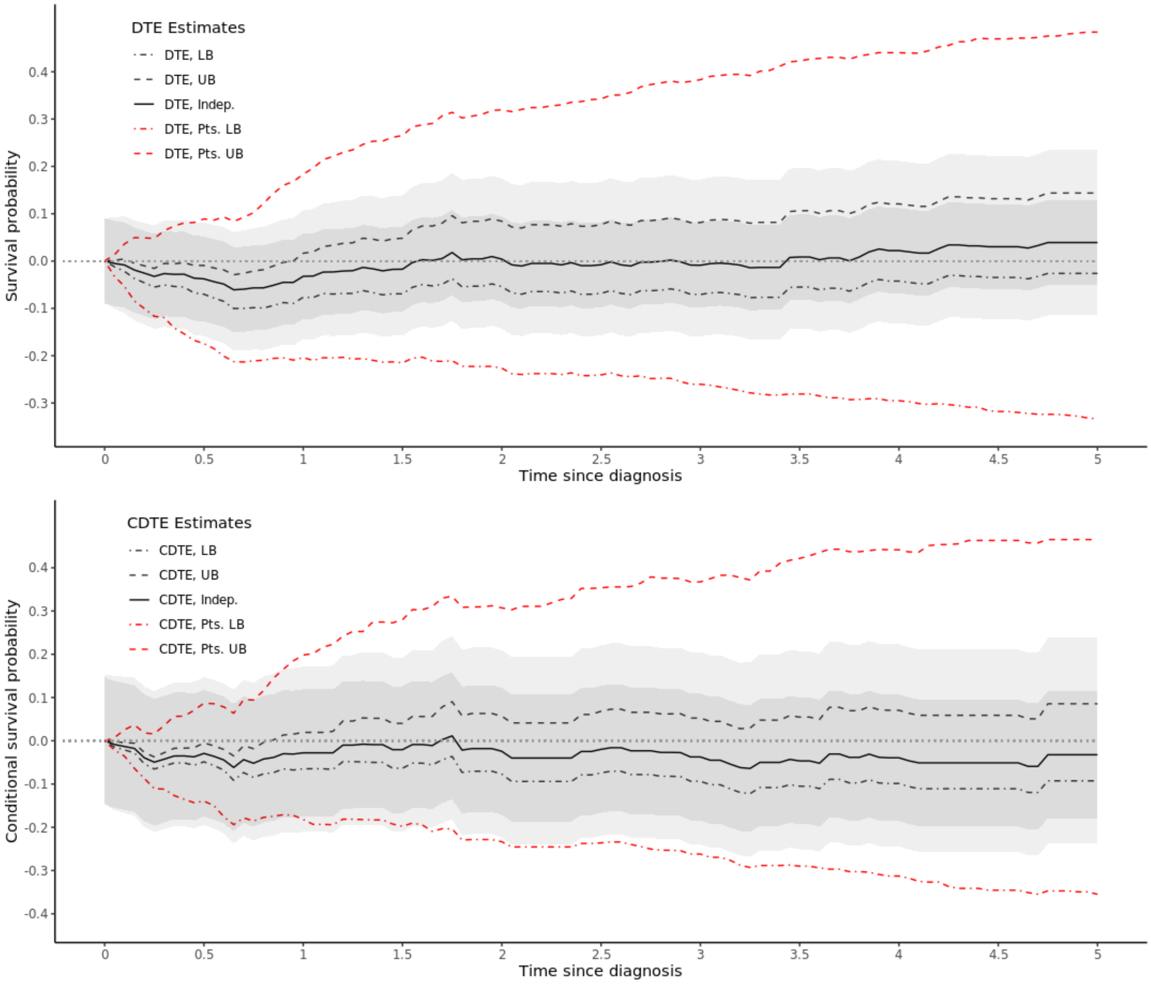
Estimation of the BGFs and TEBFs closely follows the procedures described in Section 1.6. As in Bernasconi et al. (2022), we consider two different time frames: 3 and 5 years post-treatment. For the shorter period, the TEBFs are estimated over the index set $\mathscr{U}_3$, which is equivalent to $\{0.05, 0.06, ..., 0.29, 0.3\}$ when the (C)QTE is considered, while $\mathscr{U}_3 = \{0, 0.1, ..., 2.9, 3.0\}$ for all other types of treatment effects. For the longer period, the index set $\mathscr{U}_5 = \{0.05, 0.06, ..., 0.44, 0.45\}$ is employed for the (C)QTE, and $\mathscr{U}_5 = \{0, 0.1, ..., 4.9, 5.0\}$ for all other types of treatment effects. For all of our analyses of the conditional treatment effect, we fix the conditioning set at the "representative" observation, which is the sample average of the baseline covariates.

Figure 1.2: Estimates of the potential EFS curves



Notes: The top plot depicts the unconditional potential survival curve estimates, whereas the bottom figure represents conditional survival curve estimates. The solid curves represent SICG estimates for the two protocols, with the independence copula. For each The shaded areas are bounded from above and below by SICG estimates, using Gumbel copula parameters of 1 and 1.5, respectively. The Peterson's worst case bounds for the treated and control group are depicted with dot-dash and dashed curves correspondingly.

Figure 1.3: Distributional treatment effect estimates



Notes: The top and bottom plots represent overall and conditional DTE estimates along with their 95% uniform confidence bands, respectively. The solid black lines and the dark gray area depict the DTE estimates and their uniform confidence bands under the independent censoring mechanism. Dashed (dot-dash) lines and the light gray area depict the upper (lower) bound of the DTE and the corresponding uniform confidence bands, with a Gumbel copula and $\boldsymbol{\theta} = (1, 1.5)$. The confidence bands are computed following the bootstrap procedures in Algorithms 1.5.1 and 1.9.1, respectively, with 1,999 bootstrap replications. The Peterson's worst case bounds for the DTE are delineated with dot-dash and dashed curves correspondingly.

We turn now to a discussion of the estimation results. Figure 1.2 presents the estimated potential EFS curves. Our findings mirror the original results of Bernasconi et al. (2022), revealing that the new protocol improves survival prospects in the initial years following treatment. However, this beneficial effect appears to taper off over a year earlier than previously indicated. Moreover, if we loosen the independent censoring condition, the beneficial effect may completely vanish. This is evidenced by the overlapping of the estimated identified sets of potential EFS, even under the stricter configuration, $\boldsymbol{\theta}_a$. It is also worth noting that our identified set is significantly narrower than the one derived from the no-information bounds, emphasizing our ability to provide a flexible middle ground compared to the most robust approach.

Table 1.3: Estimation results for conditional and overall TEBFs

(a)	Treatment Effect Estimators under Independent Censoring							
	ATE(3)	mad.DTE	mad.QTE	mad.CHTE	CATE(3)	mad.CDTE	mad.CQTE	mad.CCHTE
	0.078	-0.062	0.630	-0.071	0.051	-0.061	0.370	-0.071
	[-0.157, 0.314]	[-0.189, 0.139]	[-1.61, 2.15]	[-0.244, 0.189]	[-0.082, 0.184]	[-0.139, 0.097]	[-0.784, 1.034]	[-0.187, 0.142]
	Treatment Effect Estimators with $\boldsymbol{\theta}_a = (1,\ 1.5)$							
	ATE(3)	mad.DTE	mad.QTE	mad.CHTE	CATE(3)	mad.CDTE	mad.CQTE	mad.CCHTE
Lower Bd.	-0.097	-0.094	-0.490	-0.145	-0.128	-0.100	-0.460	-0.121
	[-0.329, 0.136]	[-0.222, 0.128]	[-1.564, 1.324]	[-0.325, 0.181]	[-0.26, 0.004]	[-0.177, 0.076]	[-1.038, 0.728]	[-0.24, 0.119]
Upper Bd.	0.198	0.090	0.800	0.129	0.193	0.096	0.680	0.152
	[-0.034, 0.431]	[-0.168, 0.221]	[-1.377, 2.247]	[-0.235, 0.323]	[0.059, 0.327]	[-0.108, 0.175]	[-0.512, 1.232]	[-0.16, 0.277]
	Treatment Effect Estimators with $\boldsymbol{\theta}_b = (1,\ 2)$							
	ATE(3)	mad.DTE	mad.QTE	mad.CHTE	CATE(3)	mad.CDTE	mad.CQTE	mad.CCHTE
Lower Bd.	-0.205	-0.124	-0.720	-0.196	-0.230	-0.124	-0.680	-0.174
	[-0.442, 0.032]	[-0.251, 0.127]	[-1.648, 1.138]	[-0.386, 0.19]	[-0.366 , -0.095]	[-0.202, 0.078]	[-1.226, 0.606]	[-0.296, 0.122]
Upper Bd.	0.268	0.140	1.040	0.208	0.270	0.141	0.830	0.245
	[0.034, 0.502]	[-0.165, 0.274]	[-1.329, 2.439]	[-0.255, 0.429]	[0.134, 0.406]	[-0.091, 0.222]	[-0.451, 1.331]	[-0.152, 0.386]

(b)	Treatment Effect Estimators under Independent Censoring							
	ATE(5)	mad.DTE	mad.QTE	mad.CHTE	CATE(5)	mad.CDTE	mad.CQTE	mad.CCHTE
	0.169	-0.064	2.860	-0.099	0.018	-0.061	-1.310	0.073
	[-0.26, 0.598]	[-0.211, 0.158]	[-6.271, 9.041]	[-0.332, 0.248]	[-0.227, 0.263]	[-0.151, 0.129]	[-4.084, 3.144]	[-0.228, 0.231]
	Treatment Effect Estimators with $\boldsymbol{\theta}_a = (1,\ 1.5)$							
	ATE(5)	mad.DTE	mad.QTE	mad.CHTE	CATE(5)	mad.CDTE	mad.CQTE	mad.CCHTE
Lower Bd.	-0.219	-0.122	-2.320	-0.211	-0.359	-0.100	-2.290	-0.133
	[-0.651, 0.213]	[-0.268, 0.146]	[-7.76, 5.69]	[-0.451, 0.24]	[-0.596 , -0.123]	[-0.189, 0.088]	[-5.017, 2.877]	[-0.298, 0.164]
Upper Bd.	0.407	0.090	3.550	0.163	0.288	0.144	1.060	0.299
	[-0.027, 0.841]	[-0.189, 0.242]	[-4.023, 7.643]	[-0.317, 0.439]	[0.048, 0.529]	[-0.12, 0.234]	[-1.185, 2.285]	[-0.222, 0.486]
	Treatment Effect Estimators with $\boldsymbol{\theta}_b = (1,\ 2)$							
	ATE(5)	mad.DTE	mad.QTE	mad.CHTE	CATE(5)	mad.CDTE	mad.CQTE	mad.CCHTE
Lower Bd.	-0.458	-0.152	-3.220	-0.273	-0.570	-0.124	-2.830	-0.192
	[-0.877 , -0.039]	[-0.296, 0.144]	[-8.706, 5.696]	[-0.524, 0.252]	[-0.815 , -0.325]	[-0.21, 0.087]	[-5.504, 2.734]	[-0.356, 0.164]
Upper Bd.	0.537	0.157	4.090	0.323	0.427	0.199	1.430	0.442
	[0.108, 0.966]	[-0.183, 0.309]	[-3.954, 8.114]	[-0.355, 0.644]	[0.187, 0.668]	[0.107, 0.291]	[-1.097, 2.577]	[0.241, 0.644]

Notes: The results in Panel (a) and (b) are generated using the index sets $\mathscr{U}_3$ and $\mathscr{U}_5$, respectively. Estimates of treatment effects are displayed in the first row of each panel. Except for the ATE and CATE, the value with the maximum absolute deviation (mad) from 0 over the index set is reported for each treatment effect. Numbers in square brackets represent the corresponding 95% bootstrap confidence intervals, based on 1,999 bootstrap replications. These are calculated following Algorithms 1.5.1 and 1.9.1, for the overall and conditional cases, respectively.

The findings are further validated by the overall and conditional DTE estimates illustrated in Figure 1.3. According to the uniform confidence bands of (C)DTE under independent censoring, the newer protocol does not statistically significantly outperform the older one, even in the shorter post-treatment period. The introduction of dependent censoring does not alter this conclusion. That said, this conclusion is valid only under the maintained ranges of θ, which include the special case of non-informative censoring. In order to generalize this to other dependence patterns, we would need to perform robustness checks with corresponding levels of θ.

More comprehensive results are compiled in Table 1.3. Here, we not only report the “mean” values for (C)DTE as captured by (C)ATE, but also provide the values with the maximum absolute deviations from zero over their index sets for other types of TEBFs. Additionally, the table includes 95% uniform confidence sets corresponding to these reported TEBF estimates. Although Table 1.3 does not conclusively demonstrate treatment effect nullity across the entire Gumbel copula family, it does suggest that there are no significant differences between the two protocols, on average and uniformly over the index sets, regardless of the type of policy effect under consideration. This observation aligns with the findings depicted in Figures 1.2 and 1.3. Furthermore, this conclusion remains valid across different correlation scenarios and analysis periods.

In sum, when we deviate from the conditional independence censoring mechanism, we do not find enough evidence supporting that AHOPCA ALL-2008 leads to more favorable early-year survival prospects.

1.8 Conclusion

In this paper, we proposed a framework for conducting sensitivity analysis on various treatment effect parameters when the latent duration is subject to dependent censoring. In order to obtain bounds of policy effects, we first derived bounds on the distribution of the potential outcome. Such bounds follow naturally from the concordance ordering we imposed on the Archimedean copulas. Moreover, we embedded a single-index structure into our identification framework, as an attempt to curb the "curse of dimensionality" and to make our method practically feasible. Given these results, we then proposed estimation procedures and established asymptotic properties of the resulting semiparametric estimators.

To conduct uniformly valid inference, we proposed multiplier bootstrap procedures that are easy to implement, and showed the uniform confidence sets thus constructed are asymptotically accurate. Monte Carlo simulations confirm our theoretical findings. Applying our methodology to real data, we revisited Bernasconi et al. (2022). Under a conditional independence assumption, and when the early-year survival prospect is concerned, they conclude in favor of AHOPCA ALL-200. Our sensitivity analysis demonstrates that such a conclusion may not continue to hold when we depart from the assumption of random censoring.

While we limit our discussion to the classical unconfoundedness design in this paper, our proposed methodology can be applied to other policy setups, such as the local treatment effect framework, as considered by Imbens and Angrist (1994), Angrist et al. (1996), Abadie (2003) and Frölich and Melly (2013); difference-in-differences models, cf. Card and Krueger (1994), Heckman et al. (1997), Abadie (2005), and Athey and Imbens (2006); and marginal treatment effect setup, such as Heckman and Vytlacil (2001), and Heckman and Vytlacil (2005). Upon appropriate modifications of the identification assumptions, BGFs and TEBFs can be easily derived along similar lines of Theorem 1.2 and Proposition 1.1, based on which, many policy relevant questions can be addressed.

1.9 Supplementary Appendix

This supplemental appendix contains proofs of the main theorems, auxiliary lemmas, and results. Section 1.9.1 collects the proofs of the main results of the paper. Section 1.9.2 introduce additional results on the single-index estimator, and Section 1.9.3 presents auxiliary results.

1.9.1 Proofs of Main Results

1.9.1.1 Proof of Results from Section 1.3

1.9.1.1.1 Identification of Single-index Parameters

Proof of Lemma 1.1. The proof is based on the so-called decomposition and contraction relationships of the Graphoid axioms by Dawid (1979):

$$\text{Decomposition: } A \perp\!\!\!\perp (B,C)|D \text{ implies that } A \perp\!\!\!\perp B|D,$$
$$\text{Contraction: } A \perp\!\!\!\perp B|D \text{ and } A \perp\!\!\!\perp C|(B,D) \text{ implies that } A \perp\!\!\!\perp (B,C)|D,$$

for generic random variables A,B,C,D. For each $d=0,1$, $(T_d,C_d) \perp\!\!\!\perp D|X$, under Assumption 1.1. Since the sigma field generated by $X\gamma_d$ is a subset of that generated by X, we have $(T_d,C_d) \perp\!\!\!\perp D|X,X\gamma_d$. Together with the index sufficiency condition, $(T_d,C_d) \perp\!\!\!\perp X|X\gamma_d$, we deduce from the contraction relationship that $(T_d,C_d) \perp\!\!\!\perp (D,X)|X\gamma_d$, which implies that $(T_d,C_d) \perp\!\!\!\perp D|X\gamma_d$ by the decomposition relationship. Since Y_d and R_d are deterministic functions of (T_d,C_d), the desired result follows. ■

Proof of theorem 1.1. For the first half of part (i), note that for any $(t,x,d,r) \in \mathscr{T} \times \mathscr{X} \times \{0,1\}^2$,

$$\mathbb{E}[\mathbb{1}\{D=d,R=r,Y\leq t\}|X] = F_{Y,R|D,X}(t,r|d,X)\,\mathbb{E}[\mathbb{1}\{D=d\}|X],$$

where $F_{Y,R|D,X}(t|d,X) = F_{Y_d,R_d|D,X}(t,r|d,X) = F_{Y_d,R_d|X}(t,r|X) = F_{Y_d,R_d|X\gamma_d}(t,r|X\gamma_d)$, almost surely. The second equality follows under Assumption 1.1, and the last holds under Assumption 1.2. On the other hand,

$$\mathbb{E}[G_{d,r}(t,X\gamma_d)|X] = G_{d,r}(t,X\gamma_d) = F_{Y_d,R_d|D,X\gamma_d}(t,r|d,X\gamma_d) = F_{Y_d,R_d|X\gamma_d}(t,r|X\gamma_d),$$

almost surely. The third equality is due to Lemma 1.1. Thus, (1.3.3) holds almost surely.

The converse part can be established by contradiction, applying similar arguments as in the proof of Theorem 4.1 in Ichimura (1993). Suppose there exists $d\in\{0,1\}$ and $\gamma^*\in\Gamma$, such that $\gamma^*\neq\gamma_d$, and $\mathbb{E}[U_{\gamma^*}(t,d,r)|X]=0$ almost everywhere for $(t,r)\in\mathscr{T}_0\times\{0,1\}^2$. Note that under Assumption 1.2, it continues to hold that $\mathbb{E}[\mathbb{1}\{D=d,R=r,Y\leq t\}|X]$ $= F_{Y_d,R_d|X\gamma_d}(t,r|X\gamma_d)\,\mathbb{E}[\mathbb{1}\{D=d\}|X]$ almost surely.

For any $x\in\mathscr{X}_0$, let $v=x\gamma^*$ and $\bar{\gamma}=\gamma_d-\gamma^*$, we have

$$F_{Y_d,R_d|X\gamma_d}(t,r|x\gamma_d) = F_{Y_d,R_d|X\gamma_d}\left(t,r|v+\sum_{\ell=1}^{k-1}\bar{\gamma}_\ell x_{[\ell+1]}\right) = G_{d,r}(t,v),$$

where the second equality holds almost surely. Fix v and take partial derivative of the middle term in the above display with respect to $\{x_{[\ell]}\}_{\ell=2}^{k_1}$. It follows that $\partial F_{Y_d,R_d|X\gamma_d}(t,r|x\gamma)/\partial x\gamma|_{x\gamma=x\gamma_d}\cdot\bar{\gamma}_{[\ell]}=0$, for $\ell=1,...,k_1-1$. Recall our assumption on $\mathscr{X}_0$, $\partial F_{Y_d,R_d|X\gamma_d}(t,r|x\gamma)/\partial x\gamma|_{x\gamma=X\gamma_d}\neq 0$ with positive probability. Consequently, $\gamma^*_\ell=\gamma_{d,\ell}$, for $\ell=1,...,k_1-1$.

Under Assumption 1.3.3 with $\gamma=\gamma_d$, there exists an open interval $\mathscr{V}_0$ such that for all $v\in\mathscr{V}_0$,

$$F_{Y_d,R_d|X\gamma_d}(t,r|v)=F_{Y_d,R_d|X\gamma_d}(t,r|v+\bar{\gamma}_\ell)=G_{d,r}(t,v),$$

where $\ell=k_1,...,k-1$. In view of the first equality, we find from Assumption 1.3.4 (ii) that $\bar{\gamma}_\ell=0$, and thus, $\gamma^*_\ell=\gamma_{d,\ell}$, for $\ell=k_1,...,k-1$. This contradicts the supposition that $\gamma^*\neq\gamma_d$. Hence, part (ii) follows.

To show part (iii), we first note that $\mathscr{J}_d(\gamma;\vartheta)\geq$ due to its construction. Next, when $\gamma\neq\gamma_d$,

$$\mathscr{J}_d(\gamma;\vartheta)\geq\int_{\mathscr{T}_0\times\{0,1\}}\int_{z\in\mathscr{Z}}\left\|\mathbb{E}[U_{d,\gamma,\ell}(t,r)\vartheta(X;z)]\right\|^2 d\Pi_Z(z)d\Pi_{T,R}(t,r)>0$$

where the second inequality follows because ϑ is chosen such that the equivalence in (1.3.1) holds, and from part (ii), we have $\mathbb{E}[U_{d,\gamma,\ell}(t,r)|X]>0$, $\forall(d,r,t)\in\{0,1\}^2\times\mathscr{T}_0$. ■

1.9.1.1.2 Identification of Treatment Effects

Proof of Theorem 1.2. Proof of the first half of part (i) is a slight modification of Lemma 1 of Braekers and Veraverbeke (2005). For a fixed $x\in\mathscr{X}$, we know from Assumption 1.4.2 that the true copula is Archimedean and indexed by $\theta^*_d(x)$, and

$$\begin{aligned}\partial S_{Y_d,R|X}(\tilde{y},1|x)/\partial\tilde{y}|_{\tilde{y}=y}&=\partial S_{T_d,C_d|X}(\tilde{y},\tilde{y}|x)/\partial\tilde{y}|_{\tilde{y}=y}=\partial\mathscr{C}(S_{T_d|X}(\tilde{y}|x),S_{C_d|X}(\tilde{y}|x);\theta^*_d(x))/\partial\tilde{y}|_{\tilde{y}=y}\\&=-\frac{\phi'_{\theta^*_d(x)}(S_{T_d|X}(y|x))S'_{T_d|X}(y|x)}{\phi'_{\theta^*_d(x)}(S_{Y_d|X}(y|x))}.\end{aligned}$$

The first equality is due to Tsiatis (1975), the second is by Assumption 1.4.2, and the last is due to the construction of the Archimedean copula. Multiplying the far left and right sides by $\phi'_{\theta^*_d(x)}(S_{Y_d|X}(y|x))$ and integrating them with respect to y over $[0,t]$ gives

$$\int_0^t\phi'_{\theta^*_d(x)}(S_{Y_d|X}(y|x))S_{Y_d,R|X}(dy,1|x)=\phi_{\theta^*_d(x)}\left(S_{T_d}(t|x)\right).\qquad(1.9.1)$$

Note that

$$\begin{aligned}
s_d(y,x\gamma_d) &= \mathbb{E}\left[\mathbb{E}[D\mathbb{1}\{Y>y\}|D=d,X]|D=d,X\gamma_d=x\gamma_d\right]\\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{Y_d>y\}|X]|D=d,X\gamma_d=x\gamma_d\right]\\
&= \mathbb{E}\left[S_{Y_d|X}(y|X)|D=d,X\gamma_d=x\gamma_d\right]\\
&= \mathbb{E}\left[S_{Y_d|X\gamma_d}(y|X\gamma_d)|D=d,X\gamma_d=x\gamma_d\right]\\
&= S_{Y_d|X\gamma_d}(y|x\gamma_d) = S_{Y_d|X}(y|x),
\end{aligned}$$

where the first equality follows by the tower property of conditional expectation. The second is by Assumption 1.1, and the third is due to Assumption 1.2. The last one holds under Assumption 1.2. Same lines of arguments lead to $s_{d,1}(y,x\gamma_d) = S_{Y_d,R|X}(y,1|x)$. Substituting these equations into (1.9.1) and taking the inverse of $\phi_{\theta_d^*(x)}$ on both sides, the desired result then follows by noting that $\theta_d^*(x) \in \Theta$ for all $x \in \mathscr{X}$ and $d \in \{0,1\}$. The second half of (i) follows from the preceding analysis by taking expectation of $s_{T_d}(\cdot, X\gamma_d, \theta(X))$ with respect to X.

Part (ii) follows directly from part (i) and Proposition 2 of Rivest and Wells (2001), and therefore, the proof is omitted. ■

Proof of Proposition 1.1. Under Assumption 1.1–1.3, $\boldsymbol{\gamma}$ is identified by Theorem 1.1. When $\theta_1 \leq \theta_d^*(\cdot) \leq \theta_2$, the same arguments as in Theorem 1.2 lead to $S_{T_d}(t) \in [s_{T_d}(t,\theta_2),\ s_{T_d}(t,\theta_1)]$, for $t \in \mathscr{T}$.

The identified sets for the four types of treatment effects can then be derived using the fact that each of their corresponding treatment responses respects the stochastic dominance relations of the potential survival functions. Take the restricted ATE as an example,

$$\begin{aligned}
\mathbb{E}\left[\tilde{T}_d(t)\right] &= -\int_0^t yd(1-F_{T_d}(y)) + tS_{T_d}(t)\\
&= \int_0^t (1-F_{T_d}(y))dy - t(1-F_{T_d}(t)) + tS_{T_d}(t)\\
&= \int_0^t S_{T_d}(y)dy,
\end{aligned}$$

where the second line follows from an integration by part. Consequently, $\int_{[0,t]}\left(s_{T_1}(y,\theta_2) - s_{T_0}(y,\theta_1)\right)dy \leq \mathbb{E}\left[\tilde{T}_1(t)\right] - \mathbb{E}\left[\tilde{T}_0(t)\right] \leq \int_{[0,t]}\left(s_{T_1}(y,\theta_1) - s_{T_0}(y,\theta_2)\right)dy$.

Sharpness is inherited from that of the potential causal curves. Results for the conditional treatment effects can be shown with similar arguments. ■

1.9.1.2 Proof of Results from Section 1.4

1.9.1.2.1 Uniform Linear Representation

Proof of Theorem 1.3. The proof is similar to that of Theorem 4.1 in Fan and Liu (2018), with substantial differences due to the use of single-index estimator. We first provide a uniform linear expansion for $\phi_\theta\left(\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta)\right) - \phi_\theta\left(s_{T_d}(t,x\gamma_d,\theta)\right)$, the desired result then follows by a second order Taylor expansion of ϕ_θ^{-1}. We let $s_{d,r} \equiv 1/2 - G_{d,r}$ and its estimator $\hat{s}_{d,r} \equiv 1/2 - \hat{G}_{d,r}$ for $r = 0,1$. Observe that

$$
\begin{align}
&\phi_\theta\left(\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta)\right) - \phi_\theta\left(s_{T_d}(t,x\gamma_d,\theta)\right) \tag{1.9.2}\\
=&\int_0^t \phi_\theta'\left(\hat{s}_d(y,x\hat{\gamma}_d)\right)\hat{s}_{d,1}(dy,x\hat{\gamma}_d) - \int_0^t \phi_\theta'\left(s_d(y,x\gamma_d)\right)s_{d,1}(dy,x\gamma_d) \notag\\
=&\int_0^t \left\{\phi_\theta'\left(\hat{s}_d(y,x\hat{\gamma}_d)\right) - \phi_\theta'\left(s_d(y,x\gamma_d)\right)\right\}s_{d,1}(dy,x\gamma_d) \notag\\
&+\int_0^t \phi_\theta'\left(s_d(y,x\gamma_d)\right)\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d) - s_{d,1}(dy,x\gamma_d)\right\} \notag\\
&+\int_0^t \left\{\phi_\theta'\left(\hat{s}_d(y,x\hat{\gamma}_d)\right) - \phi_\theta'\left(s_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d) - s_{d,1}(dy,x\gamma_d)\right\} \notag\\
=&\int_0^t \phi_\theta''\left(s_d(y,x\gamma_d)\right)\left\{\hat{s}_d(y,x\hat{\gamma}_d) - s_d(y,x\gamma_d)\right\}s_{d,1}(dy,x\gamma_d) \tag{1.9.3}\\
&+\phi_\theta'\left(s_d(t,x\gamma_d)\right)\left\{\hat{s}_{d,1}(t,x\hat{\gamma}_d) - s_{d,1}(t,x\gamma_d)\right\} \tag{1.9.4}\\
&-\int_0^t \phi_\theta''\left(s_d(y,x\gamma_d)\right)\left\{\hat{s}_{d,1}(y,x\hat{\gamma}_d) - s_{d,1}(y,x\gamma_d)\right\}s_d(dy,x\gamma_d) \tag{1.9.5}\\
&+r_{n,1}(t,x,\theta) + r_{n,2}(t,x,\theta) \notag
\end{align}
$$

where

$$
\begin{aligned}
r_{n,1}(t,x,\theta) &= \frac{1}{2}\int_0^t \phi_\theta^{'''}\left(\zeta(y,x)\right)\left\{\hat{s}_d(y,x\hat{\gamma}_d) - s_d(y,x\gamma_d)\right\}^2 s_{d,1}(dy,x\gamma_d),\\
r_{n,2}(t,x,\theta) &= \int_0^t \left\{\phi_\theta'\left(\hat{s}_d(y,x\hat{\gamma}_d)\right) - \phi_\theta'\left(s_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d) - s_{d,1}(dy,x\gamma_d)\right\}.
\end{aligned}
$$

The random function ζ lies between $\hat{s}_d$ and s_d. The second equality follows by direct manipulation. The fourth line is due to a second order Taylor expansion of $\phi_\theta'\left(\hat{s}_d\right)$ around s_d, which also produces the remainder $r_{n,1}$, and the fifth one follows by an integration by part on the term in the third line.

The proof proceed in two steps: we first derive the dominating terms of (1.9.3) - (1.9.5), and then we show th two remainder terms $r_{n,1}$ and $r_{n,2}$ are asymptotically negligible.

Step 1: expansion of first-order terms.

It suffices to show (1.9.3). (1.9.4) and (1.9.5) can be handled analogously. Let $\ddot{\phi}_{d,\gamma}^{\theta}(y,x) \equiv \phi_\theta''\left(s_d(y,x\gamma)\right)$. A second order Taylor expansion of $\hat{s}_d$ with respect to γ around γ_d yields

$$\begin{aligned}
&\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{\hat{s}_d(y,x\hat{\gamma}_d)-s_d(y,x\gamma_d)\right\}s_{d,1}(dy,x\gamma_d)\\
&=\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)(\hat{s}_d(y,x\gamma_d)-s_d(y,x\gamma_d))s_{d,1}(dy,x\gamma_d)\\
&\quad-(\hat{\gamma}_d-\gamma_d)'\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\partial_\gamma\hat{G}_d(y,x\gamma_d)s_{d,1}(dy,x\gamma_d)\\
&\quad+(\hat{\gamma}_d-\gamma_d)'\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{\partial_\gamma\hat{G}_d(y,x\tilde{\gamma}_d)-\partial_\gamma\hat{G}_d(y,x\gamma_d)\right\}s_{d,1}(dy,x\gamma_d)\\
&\equiv L_{n,1}+L_{n,21}+L_{n,22},
\end{aligned}$$

where $\tilde{\gamma}_d$ lies between $\hat{\gamma}_d$ and γ_d. We rewrite the first term as

$$\begin{aligned}
L_{n,1}&=-\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\frac{\hat{\kappa}_{d,y}(x\gamma_d)}{\hat{f}_d(x\gamma_d)}-G_d(y,x\gamma_d)\right)s_{d,1}(dy,x\gamma_d)\\
&=-\frac{1}{n\hat{f}_d(x\gamma_d)}\sum_{i=1}^n K_h(x\gamma_d,X_i\gamma_d)\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\mathscr{E}_{d,\gamma_d,i}(y,x)s_{d,1}(dy,x\gamma_d)\\
&=-\frac{1}{nf_d(x\gamma_d)}\sum_{i=1}^n K_h(x\gamma_d,X_i\gamma_d)\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\mathscr{E}_{d,\gamma_d,i}(y,x)s_{d,1}(dy,x\gamma_d) &(1.9.6)\\
&\quad+\frac{\hat{f}_d(x\gamma_d)-f_d(x\gamma_d)}{nf_d(x\gamma_d)\hat{f}_d(x\gamma_d)}\sum_{i=1}^n K_h(x\gamma_d,X_i\gamma_d)\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\mathscr{E}_{d,\gamma_d,i}(y,x)s_{d,1}(dy,x\gamma_d). &(1.9.7)
\end{aligned}$$

where $\mathscr{E}_{d,\gamma,\ell}(y,x)\equiv\mathbb{1}\left\{D_\ell=d\right\}(\mathbb{1}\left\{Y_\ell\le y\right\}-G_d(y,x\gamma))$. Note that the difference between $\mathscr{E}_{d,\gamma,\ell}(y,x)$ and $\mathscr{E}_{d,\gamma,\ell}(y)$ lies in whether X is fixed at x.

We divide (1.9.6) into two parts,

$$\begin{aligned}
&-\frac{1}{nf_d(x\gamma_d)}\sum_{i=1}^n K_h(x\gamma_d,X_i\gamma_d)\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\mathscr{E}_{d,\gamma_d,i}(y)s_{d,1}(dy,x\gamma_d), &(1.9.8)\\
&-\frac{1}{nf_d(x\gamma_d)}\sum_{i=1}^n K_h(x\gamma_d,X_i\gamma_d)\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)(\mathscr{E}_{d,\gamma_d,i}(y,x)-\mathscr{E}_{d,\gamma_d,i}(y))s_{d,1}(dy,x\gamma_d) &(1.9.9)
\end{aligned}$$

The first term in the preceding display is centered and belongs to $\eta_{s,d}$. The second term corresponds to the first-order bias and is part of $\eta_{b,d}$.

By the definition of $\hat{G}_d$, (1.9.7) is equal to

$$\begin{aligned}
&\frac{\hat{f}_d(x\gamma_d)-f_d(x\gamma_d)}{f_d(x\gamma_d)}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\hat{G}_d(y,x\gamma_d)-G_d(y,x\gamma_d)\right)s_{d,1}(dy,x\gamma_d)\\
&\lesssim \sup_{\tilde{\mathscr{T}}\times\mathscr{X}}\left|\hat{f}_d(x\gamma_d)-f_d(x\gamma_d)\right|\cdot\sup_{\tilde{\mathscr{T}}\times\mathscr{X}}\left|\hat{G}_d(y,x\gamma_d)-G_d(y,x\gamma_d)\right|=O_p\left(\frac{\log n}{nh}\right),
\end{aligned}$$

uniformly over Θ. The inequality follows by (1.9.36) and (1.9.37). The equality is due to Assumption 1.9.2.

Regarding $L_{n,21}$, we have

$$
\begin{aligned}
&-(\hat{\gamma}_d-\gamma_d)'\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)G_d^{(1)}(y,x\gamma_d)s_{d,1}(dy,x\gamma_d)\\
&-(\hat{\gamma}_d-\gamma_d)'\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{\partial_\gamma\hat{G}_d(y,x\gamma_d)-G_d^{(1)}(y,x\gamma_d)\right\}s_{d,1}(dy,x\gamma_d)
\end{aligned}
$$

Under Assumption 1.6.4, $\phi''_{\theta}(u)$ is bounded on $[\upsilon_o,1]$ uniformly in θ. Meanwhile, $s_d(y,x)$ is bounded in the same interval whenever $y\in\tilde{\mathscr{T}}$, uniformly in $x\in\mathscr{X}$. Consequently, $\ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)$ is bounded on $\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta$. By (1.9.32), the last term is bounded from above by $\sup_{\tilde{\mathscr{T}}\times\mathscr{X}}\left\|\partial_\gamma\hat{G}_d(t,x\gamma_d)-G_d^{(1)}(t,x\gamma_d)\right\|\|\hat{\gamma}_d-\gamma_d\|=\left(O_p\left((\log n)^{1/2}n^{-1/2}h^{-3/2}\right)+O(h^s)\right)\cdot O_p\left(n^{-1/2}\right)$, which is $O_p\left((\log n)^{1/2}n^{-1}h^{-3/2}\right)$ under our rate condition on the bandwidth. Therefore, $L_{n,21}$ is dominated by

$$
-\frac{1}{n}\sum_{i=1}^{n}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)G_d^{(1)}(y,x\gamma_d)'s_{d,1}(dy,x\gamma_d)V_d^{-1}\psi_2(X_i,\gamma_d).
$$

From Lemma 1.8, we deduce that $L_{n,22}$ has a uniform rate of $O_p\left(n^{1/2}\right)\cdot O_p\left((\log n)^{1/2}n^{-1}h^{-5/2}\right)=o_p((\log n)^{1/2}\cdot n^{-1}h^{-3/2})$.

So far, we have derived the leading terms of (1.9.3). The other two terms, (1.9.4) and (1.9.5), can be treated analogously.

Step 2: uniform asymptotic negligibility of $r_{n,1}$ and $r_{n,2}$.

By the mean value theorem, we have

$$
\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta} r_{n,1}(t,x,\theta)\lesssim \sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|\phi'''_{\theta}\left(\zeta(t,x)\right)\right|\left\{\hat{s}_d(t,x)-s_d(t,x)\right\}^2,
$$

We can deduce from (1.9.36), and Assumption 1.6.4 that the right hand side is of order $O_p\left(\log n\cdot n^{-1}h^{-1}\right)+O(h^{2s})$, which is $O_p\left(\log n\cdot n^{-1}h^{-1}\right)$ under Assumption 1.9.2.

Next, we show $r_{n,2}=O_p\left(\log n\cdot n^{-1}h^{-1}\right)$ as well. Perform a third order Taylor expansion of ϕ'_{θ} with respect to s_d,

$$
\begin{aligned}
r_{n,2}(t,x,\theta)=&\int_0^t\left\{\ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\hat{s}_d(y,x\hat{\gamma}_d)-s_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d)-s_{d,1}(dy,x\gamma_d)\right\}\\
&+\frac{1}{2}\int_0^t\phi'''_{\theta}\left(\zeta(y,x)\right)\left\{\hat{s}_d(y,x\hat{\gamma}_d)-s_d(y,x\gamma_d)\right\}^2\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d)-s_{d,1}(dy,x\gamma_d)\right\}.
\end{aligned}
\tag{1.9.10}
$$

The second term is asymptotically dominated in view of Lemma 1.4. Focusing on the first term, we have

$$
\begin{aligned}
(1.9.10) = &\int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\hat{s}_d(y,x\hat{\gamma}_d) - \hat{s}_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d) - \hat{s}_{d,1}(dy,x\gamma_d)\right\} && (1.9.11)\\
&+ \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\hat{s}_d(y,x\hat{\gamma}_d) - \hat{s}_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\gamma_d) - s_{d,1}(dy,x\gamma_d)\right\} && (1.9.12)\\
&+ \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\hat{s}_d(y,x\gamma_d) - s_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\hat{\gamma}_d) - \hat{s}_{d,1}(dy,x\gamma_d)\right\} && (1.9.13)\\
&+ \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left(\hat{s}_d(y,x\gamma_d) - s_d(y,x\gamma_d)\right)\right\}\left\{\hat{s}_{d,1}(dy,x\gamma_d) - s_{d,1}(dy,x\gamma_d)\right\} && (1.9.14)
\end{aligned}
$$

We analyze (1.9.11), (1.9.12), and (1.9.14) in turn. Via integration by parts, (1.9.13) can be handled in the same fashion as (1.9.12), and therefore, the proof is omitted.

We provide results for (1.9.11). By a first-order Taylor expansion of $\hat{s}_d(y,x\hat{\gamma}_d)$ in γ around γ_d, we get

$$
(\hat{\gamma}_d - \gamma_d)' \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\partial_{\gamma}\hat{G}_d(y,x\tilde{\gamma}_d)\partial_{\gamma'}\hat{G}_d(dy,x\breve{\gamma}_d)(\hat{\gamma}_d - \gamma_d),
$$

where $\tilde{\gamma}_d$ and $\breve{\gamma}_d$ lie between $\hat{\gamma}_d$ and γ_d. Let $I_{d,t,1,i} = R_i \mathbb{1}\{D_i = d, Y_i \leq t\}$. Expanding the partial derivative in the integrator, we get

$$
\begin{aligned}
&\frac{(\hat{\gamma}_d - \gamma_d)'}{nh^2\hat{f}_d(x\breve{\gamma}_d)} \sum_{i=1}^n I_{d,t,1,i}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)\partial_{\gamma}\hat{G}_d(Y_i,x\tilde{\gamma}_d)K^{(1)}\left((X_i\breve{\gamma}_d - x\breve{\gamma}_d)/h\right)(X_{[-1],i} - x_{[-1]})'(\hat{\gamma}_d - \gamma_d)\\
&- \frac{(\hat{\gamma}_d - \gamma_d)'}{nh\hat{f}_d(x\breve{\gamma}_d)^2} \sum_{i=1}^n I_{d,t,1,i}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)\partial_{\gamma}\hat{G}_d(Y_i,x\tilde{\gamma}_d)K((X_i\breve{\gamma}_d - x\breve{\gamma}_d)/h)\partial_{\gamma'}\hat{f}_d(x\breve{\gamma}_d)(\hat{\gamma}_d - \gamma_d)\\
&\equiv L_{n,31} + L_{n,32}.
\end{aligned}
$$

Rewrite $L_{n,31}$ as

$$
\begin{aligned}
&\frac{(\hat{\gamma}_d - \gamma_d)'}{nh^2 f_d(x\breve{\gamma}_d)} \sum_{i=1}^n I_{d,t,1,i}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)\partial_{\gamma}\hat{G}_d(Y_i,x\tilde{\gamma}_d)K^{(1)}\left((X_i\breve{\gamma}_d - x\breve{\gamma}_d)/h\right)(X_{[-1],i} - x_{[-1]})'(\hat{\gamma}_d - \gamma_d) && (1.9.15)\\
&- \frac{(\hat{\gamma}_d - \gamma_d)'(\hat{f}_d(x\breve{\gamma}_d) - f_d(x\breve{\gamma}_d))}{nh^2 f_d(x\breve{\gamma}_d)\hat{f}_d(x\breve{\gamma}_d)}\\
&\cdot \sum_{i=1}^n I_{d,t,1,i}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)\partial_{\gamma}\hat{G}_d(Y_i,x\tilde{\gamma}_d)K^{(1)}\left((X_i\breve{\gamma}_d - x\breve{\gamma}_d)/h\right)(X_{[-1],i} - x_{[-1]})'(\hat{\gamma}_d - \gamma_d). && (1.9.16)
\end{aligned}
$$

The term in (1.9.15) can be further decomposed as

$$
\frac{(\hat{\gamma}_d - \gamma_d)'}{nhf_d(x\breve{\gamma}_d)} \sum_{i=1}^n I_{d,t,1,i}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)G^{(1)}_d(Y_i,x\tilde{\gamma}_d)K^{(1)}_h(x\gamma,X_i\gamma)(X_{[-1],i} - x_{[-1]})'(\hat{\gamma}_d - \gamma_d)
$$

$$- \frac{(\hat{\gamma}_d - \gamma_d)'}{nhf_d(x\breve{\gamma}_d)} \sum_{i=1}^{n} I_{d,t,1,i} \ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)(\partial_\gamma \hat{G}_d(Y_i,x\tilde{\gamma}_d) - G_d^{(1)}(Y_i,x\tilde{\gamma}_d))K_h^{(1)}(x\gamma,X_i\gamma)(X_{[-1],i} - x_{[-1]})'(\hat{\gamma}_d - \gamma_d).$$

From (1.9.33), we find that, Assumptions 1.3.1 and 1.6, $G_d^{(1)}(y,x\gamma)$ is bounded uniformly on $\tilde{\mathscr{T}} \times \mathscr{X}_\Gamma$. Additionally, $\ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)$ is bounded on $\tilde{\mathscr{T}} \times \mathscr{X} \times \Theta$ under Assumption 1.6.4. Hence, the first term can be bounded from above by

$$\frac{M_1}{h} \|\hat{\gamma}_d - \gamma_d\|^2 \sup_{(y,x,\gamma) \in \tilde{\mathscr{T}} \times \mathscr{X} \times \Gamma_{d,n}} \left\{ \left\| G_d^{(1)}(y,x\gamma) \right\| \|x_{[-1]}\| \right\} \sup_{(x,\gamma) \in \mathscr{X} \times \Gamma_{d,n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| K_h^{(1)}(x\gamma,X_i\gamma) \right| \right\}$$
$$= O_p(n^{-1}h^{-1}) = o_p\left(\log n \cdot n^{-1}h^{-1}\right),$$

where the last equality holds uniformly over Θ. In view of (1.9.37), we deduce that the second is bounded by

$$\frac{M_2}{h} \|\hat{\gamma}_d - \gamma_d\|^2 \sup_{(y,x,\gamma) \in \tilde{\mathscr{T}} \times \mathscr{X} \times \Gamma_{d,n}} \left\{ \left\| \partial_\gamma \hat{G}_d(y,x\gamma) - G_d^{(1)}(y,x\gamma) \right\| \|x_{[-1]}\| \right\} \sup_{x \in \mathscr{X}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| K_h^{(1)}(x\gamma,X_i\gamma) \right| \right\}$$
$$= O_p(n^{-1}h^{-1}) \cdot \left(O_p\left((\log n)^{1/2} n^{-1/2} h^{-3/2} \right) + O(h^s) \right) = o_p\left(\log n \cdot n^{-1}h^{-1}\right),$$

uniformly over Θ. Applying similar reasoning, we are able to show that (1.9.16) is the order $O_p((\log n)^{1/2} n^{-3/2} \cdot h^{-3/2})$, and $L_{n,32} = O_p\left(n^{-1}\right)$, both of which are $o_p\left(\log n \cdot n^{-1}h^{-1}\right)$.

Next, we bound (1.9.12). A Taylor expansion of $\hat{s}_d(y,x\hat{\gamma}_d)$ around γ_d yields,

$$(1.9.12) = -(\hat{\gamma}_d - \gamma_d)' \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \partial_\gamma \hat{G}_d(y,x\gamma_d) \right\} \left\{ \hat{s}_{d,1}(dy,x\gamma_d) - s_{d,1}(dy,x\gamma_d) \right\}$$
$$+ (\hat{\gamma}_d - \gamma_d)' \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left(\partial_\gamma \hat{G}_d(y,x\tilde{\gamma}_d) - \partial_\gamma \hat{G}_d(y,x\gamma_d) \right) \right\} \left\{ \hat{s}_{d,1}(dy,x\gamma_d) - s_{d,1}(dy,x\gamma_d) \right\}$$
$$\equiv (\hat{\gamma}_d - \gamma_d)' B_{n,1} + (\hat{\gamma}_d - \gamma_d)' B_{n,2}.$$

$B_{n,1}$ can be rewritten as

$$B_{n,1} = \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \partial_\gamma \hat{G}_d(y,x\gamma_d) \right\} d \left\{ \frac{\hat{\kappa}_{d,1,y}(x\gamma_d)}{f_d(x\gamma_d)} - G_{d,1}(y,x\gamma_d) \right\}$$
$$- \frac{\hat{f}_d(x\gamma_d) - f(x\gamma_d,d)}{f(x\gamma_d,d)\hat{f}_d(x\gamma_d)} \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \partial_\gamma \hat{G}_d(y,x\gamma_d) \right\} d\hat{\kappa}_{d,1,y}(x\gamma_d).$$

Similar analysis along the lines of $L_{n,31}$ gives a uniform bound of $O_p\left((\log n)^{1/2} n^{-1/2} h^{-1/2} \right)$ for the second term.

Expanding the partial derivative in the first term leads to

$$
\begin{aligned}
&\frac{1}{f(x\gamma_d,d)^2}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\partial_\gamma\hat{\kappa}_{d,y}(x\gamma_d)d\left[\hat{\kappa}_{d,1,y}(x\gamma_d)-f(x\gamma_{d,1},d)G_{d,1}(y,x\gamma_d)\right] && (1.9.17)\\
&-\frac{\partial_\gamma\hat{f}_d(x\gamma_d)}{f_d(x\gamma_d)\hat{f}_d(x\gamma_d)^2}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\hat{\kappa}_{d,y}(x\gamma_d)d\left[\hat{\kappa}_{d,1,y}(x\gamma_{d,1})-f_d(x\gamma_d)G_{d,1}(y,x\gamma_d)\right] && (1.9.18)\\
&-\frac{(\hat{f}_d(x\gamma_d)-f(x\gamma_d,d))}{f(x\gamma_d,d)^2\hat{f}_d(x\gamma_d)}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\partial_\gamma\hat{\kappa}_{d,y}(x\gamma_d)d\left[\hat{\kappa}_{d,1,y}(x\gamma_d)-f_d(x\gamma_d)G_{d,1}(y,x\gamma_d)\right]\\
&+\frac{\partial_\gamma\hat{f}_d(x\gamma_d)(\hat{f}_d(x\gamma_d)^2-f(x\gamma_d,d)^2)}{f(x\gamma_d,d)^3\hat{f}_d(x\gamma_d)^2}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\hat{\kappa}_{d,y}(x\gamma_d)d\left[\hat{\kappa}_{d,1,y}(x\gamma_d)-f_d(x\gamma_d)G_{d,1}(y,x\gamma_{d,1})\right].
\end{aligned}
$$

As $\hat{f}_d(x\gamma_d)$ converges uniformly to $f_d(x\gamma_d)$ in probability, the last two terms are clearly dominated by the first two in the limit. We therefore focus on the convergence of (1.9.17) and (1.9.18).

Let $\kappa_{d,y}(x\gamma)=\mathbb{E}[\mathbb{1}\{D=d,Y\leq y\}K_h(x\gamma,X\gamma)]$ and $\kappa_{d,1,y}(x\gamma)=\mathbb{E}[R\mathbb{1}\{D=d,Y\leq y\}\cdot K_h(x\gamma,X\gamma)]$. The integrator of (1.9.17) can be decomposed into a centered term $\nu_{d,1}(y,x,\gamma_d)=\big(\hat{\kappa}_{d,1,y}(x\gamma_d)-\kappa_{d,1,y}(x\gamma_d)\big)$ and a bias term $\mu_{d,1}(y,x,\gamma_d)=\big(\kappa_{d,1,y}(x\gamma_d)-f_d(x\gamma_d)G_{d,1}(y,x\gamma_d)\big)$.

Regarding the centered part, let us define

$$
\begin{aligned}
L_{n,41,\ell}\equiv&\frac{1}{nhf(x\gamma_d,d)^2}\sum_{i=1}^n\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)I_{d,y,i}(K_h^{(1)}(x\gamma_d,X_i\gamma_d)(X_{\ell,i}-x_\ell)\\
&-\mathbb{E}[I_{d,y}K_h^{(1)}(x\gamma_d,X\gamma_d)(X_\ell-x_\ell)])\nu_{d,1}(dy,x,\gamma_d),\\
L_{n,42,\ell}\equiv&\frac{1}{f(x\gamma_d,d)^2}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\mathbb{E}[I_{d,y}h^{-1}K_h^{(1)}(x\gamma_d,X\gamma_d)(X_\ell-x_\ell)]\nu_{d,1}(dy,x,\gamma_d).
\end{aligned}
$$

The first term can be represented as a degenerate second order U process indexed by ω. Specifically,

$$
L_{n,41,\ell}(\omega)=\frac{1}{n^2h^3}\left\{\sum_{i=1}^n g_{1,\ell}(W_i,W_i,\omega)+\sum_{i\neq j}^n g_{1,\ell}(W_i,W_j,\omega)\right\}\equiv L^a_{n,41,\ell}(\omega)+L^b_{n,41,\ell}(\omega),
$$

where

$$
\begin{aligned}
g_{1,\ell}(W_1,W_2,\omega)=&\frac{1}{f(x\gamma_d,d)^2}\left\{g_{11}(W_1,\omega)g_{12,\ell}(W_2,Y_1,\omega)-\int g_{11}(w_1,\omega)g_{12,\ell}(W_2,y_1,\omega)dF_W(w_1)\right\},\\
g_{11}(W_1,\omega)=&I_{d,t,1,1}hK_h(x\gamma_d,X_1\gamma_d),\\
g_{12,\ell}(W_2,y,\omega)=&\ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{I_{d,y\wedge t,2}hK_h^{(1)}(x\gamma_d,X_2\gamma_d)(X_{2,\ell}-x_\ell)\right.\\
&\left.-\int\mathbb{1}\{d_2=d,y_2\leq y\wedge t\}hK_h^{(1)}(x\gamma_d,x_2\gamma_d)(x_{2,\ell}-x_\ell)dF_W(w_2)\right\}.
\end{aligned}
$$

Direct calculation shows

$$\sup_{\omega\in\Omega}\left|L^{a}_{n,41,\ell}(\omega)\right| \lesssim \frac{1}{nh}\sup_{x\in\mathscr{X}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|K_h^{(1)}(x\gamma_d,X_i\gamma_d)K_h(x\gamma_d,X_i\gamma_d)\right|\right\} = O_p\left(\frac{1}{nh^2}\right).$$

Now, define the following class of functions

$$\mathscr{G}_1 \equiv \left\{(w_1,w_2)\mapsto g_{1,\ell}(w_1,w_2,\omega) : \ell\in\{2,...,k\},\ \omega\in\Omega\right\}. \tag{1.9.19}$$

By Lemma 1.7, it belongs to the VC type class with a bounded envelop. Standard calculations reveal that the maximum variance of U process kernel $\sup_{g\in\mathscr{G}_1}\mathbb{E}[g^2]$ is of the order $O(h^2)$. By the maximal inequality in Lemma 1.5, we conclude that $\mathbb{E}\left[\sup_{\omega\in\Omega}\left|n^{-2}\sum_{i\neq j}^{n} g_{1,\ell}(W_i,W_j,\omega)\right|\right] = O\left(\log n\cdot n^{-1}h\right)$. Applying the Markov inequality and multiplying the U statistic by h^{-3}, we deduce that $\sup_{\omega\in\Omega}\left|L^{b}_{n,41,\ell}(\omega)\right| = O_p\left(\log n\cdot n^{-1}h^{-2}\right)$.

By Lemma 1.4, the expectation in side $L_{n,42,\ell}$ is uniformly convergent to $\partial_{x\gamma}G_{d,1}(y,x\gamma_d)(x_\ell-\mathbb{E}[X_\ell|x\gamma_d])$. Thus,

$$\begin{aligned}L_{n,42,\ell} \equiv & \frac{1}{nhf(x\gamma_d,d)^2}\sum_{i=1}^{n}\Big\{I_{d,t,1,i}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x)\left(\partial_{x\gamma}G_{d,1}(Y_i,x\gamma_d)(x_\ell-\mathbb{E}[X_\ell|x\gamma_d])\right)hK_h(x\gamma_d,X_i\gamma_d)\\ & -\mathbb{E}\left[I_{d,t,1}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y,x)\left(\partial_{x\gamma}G_{d,1}(Y,x\gamma_d)(x_\ell-\mathbb{E}[X_\ell|x\gamma_d])\right)hK_h(x\gamma_d,X\gamma_d)\right]\Big\}+h^s r_{n,3}(t,x,\theta),\end{aligned}$$

where $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}|r_{n,3}(t,x,\theta)| = o_p(1)$. The first term on the right hand side can be bounded via the maximal inequality as long as the following class of function is of the VC type:

$$\mathscr{G}_2 \equiv \left\{w\mapsto g_{2,\ell}(w,\omega) : \ell\in\{2,...,k\},\ \omega\in\Omega\right\}, \tag{1.9.20}$$

where

$$\begin{aligned}& g_{2,\ell}(W,\omega) = f(x\gamma_d,d)^{-2}\Big\{I_{d,t,1}\ddot{\phi}^{\theta}_{d,\gamma_d}(Y,x)\left(\partial_{x\gamma}G_{d,1}(Y,x\gamma_d)(x_\ell-\mathbb{E}[X_\ell|x\gamma_d])\right)hK_h(x\gamma_d,X\gamma_d)\\ & -\int r_1\mathbb{1}\{d_1=d,y_1\leq t\}\ddot{\phi}^{\theta}_{d,\gamma_d}(y_1,x)\left(\partial_{x\gamma}G_{d,1}(y_1,x\gamma_d)(x_\ell-\mathbb{E}[X_\ell|x\gamma_d])\right)hK_h(x\gamma_d,x_1\gamma_d)dF_W(w_1)\Big\}.\end{aligned}$$

Note that $\sup_{g_2\in\mathscr{G}_2}\mathbb{E}[g_2^2]=O(h)$. Consequently, $L_{n,42,\ell}=O_p\left((\log n)^{1/2}\cdot n^{-1/2}h^{-1/2}\right)$ by the maximal inequality from Lemma 1.5

Turning to the bias part of (1.9.17), we define

$$L_{n,5,\ell} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{X_{\ell,i}-x_\ell}{hf(x\gamma_d,d)^2}\int_0^t\ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)I_{d,y,i}K_h^{(1)}(x\gamma_d,X_i\gamma_d)\mu_{d,1}(dy,x,\gamma_d),$$

for $\ell = 2, ..., k$. Integrating by parts gives

$$
\begin{aligned}
L_{n,5,\ell} =& \frac{1}{n}\sum_{i=1}^{n} \frac{X_{\ell,i} - x_\ell}{hf(x\gamma_d, d)^2} \ddot{\phi}^{\theta}_{d,\gamma_d}(t,x) I_{d,y,i} K_h^{(1)}(x\gamma_d, X_i\gamma_d)\mu_{d,1}(t,x,\gamma_d) \\
& - \frac{X_{\ell,i} - x_\ell}{hf(x\gamma_d, d)^2} \int_0^t I_{d,y,i} K_h^{(1)}(x\gamma_d, X_i\gamma_d)\mu_{d,1}(y,x,\gamma_d)\ddot{\phi}^{\theta}_{d,\gamma_d}(dy,x) \\
& - \frac{X_{\ell,i} - x_\ell}{hf(x\gamma_d, d)^2} I_{d,t,i}\mu_{d,1}(Y_i,x,\gamma_d)\ddot{\phi}^{\theta}_{d,\gamma_d}(Y_i,x) K_h^{(1)}(x\gamma_d, X_i\gamma_d),
\end{aligned}
\tag{1.9.21}
$$

Due to the uniform convergence of the gradient estimator by (1.9.35), the first term is uniformly bounded from above by

$$
\begin{aligned}
\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta} & \left\{ \left| f_d(x\gamma_d)^{-2} \right| \left| \ddot{\phi}^{\theta}_{d,\gamma_d}(t,x) \right| \cdot \left\{ \left\| G_d^{(1)}(t,x\gamma_d) \right\| + \left\| \partial_\gamma \hat{G}_d(t,x\gamma_d) - G_d^{(1)}(t,x\gamma_d) \right\| \right\} \right\} \\
& \cdot \sup_{(t,x)\in\tilde{\mathscr{T}}\times\mathscr{X}} \left\{ \left| \mu_{d,1}(t,x,\gamma_d) \right| \right\} = O_p(1)\cdot O(h^s).
\end{aligned}
$$

Since $\phi''(\cdot)$ and $s_d(\cdot, x\gamma_d)$ are both continuously differentiable with bounded derivative under Assumption 1.6, we can deduce from the mean value theorem that the second term is also of the order $O_p(h^s)$. Standard bias calculation yields

$$
(1.9.21) = -\frac{\int u^s K(u)du}{nh^{2-s}} \sum_{i=1}^{n} g_{3,\ell}(W_i,\omega) + h^{s+1} r_{n,4,\ell}(t,x,d,\theta),
$$

where $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta} \left| r_{n,4,\ell}(t,x,d,\theta) \right| = O_p(1)$ and, for $\ell \in \{2,...,k\}$,

$$
g_{3,\ell}(W,\omega) = \frac{X_\ell - x_\ell}{f(x\gamma_d,d)^2} I_{d,t} \ddot{\phi}^{\theta}_{d,\gamma_d}(Y,x) \partial^s_{x\gamma} \left\{ f_d(x\gamma_d) G_{d,1}(Y,x\gamma_d) \right\} hK_h^{(1)}(x\gamma_d, X\gamma_d).
$$

Once again, Lemma 1.7 establishes that

$$
\mathscr{G}_3 \equiv \left\{ w \mapsto g_{3,\ell}(w,\omega) : \ell \in \{2,...,k\},\ \omega \in \Omega \right\}, \tag{1.9.22}
$$

is of VC type with bounded envelop. Maximal variance is of the order $O(h)$. We then deduce from Lemma 1.5 and the Markov inequality that (1.9.21) is of the order $O_p\left((\log n)^{1/2} n^{-1/2} h^{s-3/2} \right)$, uniformly over $\omega \in \Omega$. Overall, (1.9.17) is $O_p\left((\log n)^{1/2} \cdot n^{-1/2} h^{-1/2} \right)$. Following same arguments, we can deduce that (1.9.18) is $O_p\left((\log n)^{1/2} \cdot n^{-1/2} h^{1/2} \right)$.

The same procedure can be followed to decompose $B_{n,2}$. In what follows, we derive the convergence rate for the

$\hat{\kappa}_{d,1,y}$ part only since the denominator $\hat{f}_d$ can be treated analogously. Specifically, define

$$\begin{aligned}
B_{n,21} \equiv & f(x\gamma_d)^{-1} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left\{ \partial_\gamma \hat{G}_d(y, x\tilde{\gamma}_d) - \partial_\gamma \hat{G}_d(y, x\gamma_d) \right\} \nu_{d,1}(dy, x, \gamma_d) \\
= & f(x\gamma_d)^{-1} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left\{ \frac{\partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d)}{\hat{f}_d(x\tilde{\gamma}_d)} - \frac{\partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)}{\hat{f}_d(x\gamma_d)} \right\} \nu_{d,1}(dy, x, \gamma_d), \\
& - f(x\gamma_d)^{-1} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left\{ \frac{\hat{\kappa}_{d,y}(x\tilde{\gamma}_d) \partial_\gamma \hat{f}_d(x\tilde{\gamma}_d)}{\hat{f}^2_d(x\tilde{\gamma}_d)} - \frac{\hat{\kappa}_{d,y}(x\gamma_d) \partial_\gamma \hat{f}_d(x\gamma_d)}{\hat{f}^2_d(x\gamma_d)} \right\} \nu_{d,1}(dy, x, \gamma_d).
\end{aligned}$$

From similar arguments applied in Lemma 1.8, one deduces that each of the two terms is dominated a degenerate second order U process that converges at a rate of $O_p\left(\log n \cdot n^{-1} h^{-7/2} \delta_n\right)$, uniformly for $\|\tilde{\gamma}_d - \gamma_d\| \leq \delta_n$. Since $\delta_n = O_p\left(n^{-1/2}\right)$, we find from Assumption 1.9.2 that $B_{n,21} = o_p\left(\log n \cdot n^{-1/2} h^{-1}\right)$.

$$\begin{aligned}
B_{n,22} \equiv & f(x\gamma_d)^{-1} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left\{ \partial_\gamma \hat{G}_d(y, x\tilde{\gamma}_d) - \partial_\gamma \hat{G}_d(y, x\gamma_d) \right\} \mu_{d,1}(dy, x, \gamma_d) \\
= & f_d(x\gamma_d)^{-2} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left\{ \partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d) - \partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d) \right\} \mu_{d,1}(dy, x, \gamma_d) \\
& - \frac{\hat{f}_d(x\tilde{\gamma}_d) - \hat{f}_d(x\gamma_d)}{f_d(x\tilde{\gamma}_d) f_d(x\gamma_d)^2} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d) \mu_{d,1}(dy, x, \gamma_d) \\
& - f(x\gamma_d)^{-1} \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left\{ \frac{\hat{\kappa}_{d,y}(x\tilde{\gamma}_d) \partial_\gamma \hat{f}_d(x\tilde{\gamma}_d)}{\hat{f}^2_d(x\tilde{\gamma}_d)} - \frac{\hat{\kappa}_{d,y}(x\gamma_d) \partial_\gamma \hat{f}_d(x\gamma_d)}{\hat{f}^2_d(x\gamma_d)} \right\} \mu_{d,1}(dy, x, \gamma_d) + (s.o.) \\
\equiv & B^a_{n,22}(t, x, \theta) + B^b_{n,22}(t, x, \theta) + B^c_{n,22}(t, x, \theta) + (s.o.)
\end{aligned}$$

Integration by parts turns $B^a_{n,22}$ into three terms. Applying arguments of Lemma 1.8, and properly accounting for the biases, to each of the terms, one deduces that $\sup_{\|\gamma - \gamma_d\| \leq \delta_n} \sup_{(t,x,\theta) \in \tilde{\mathscr{T}} \times \mathscr{X} \times \Theta} \left\| B^a_{n,22}(t, x, \theta) \right\| = O_p((\log n)^{1/2} n^{-1/2} \cdot h^{s-5/2} \delta_n)$. The same uniform rate applies to $B^b_{n,22}$, and, after further decomposition, to $B^c_{n,22}$.

Collect results on $B_{n,1}$ and $B_{n,2}$, and multiply them by $O_p\left(n^{-1/2}\right)$. We conclude that (1.9.12) is $o_p\left(\log n \cdot n^{-1} h^{-1}\right)$.

Lastly, we bound (1.9.14). Rewriting the term as

$$\begin{aligned}
& \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \left(\nu_{n,d}(y, x\gamma_d) + \mu_d(y, x\gamma_d) \right) \right\} \left\{ \nu_{n,d,1}(dy, x\gamma_d) + \mu_{d,1}(dy, x\gamma_d) \right\} \\
= & \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \nu_{n,d}(y, x\gamma_d) \right\} \nu_{n,d,1}(dy, x\gamma_d) + \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \nu_{n,d}(y, x\gamma_d) \right\} \mu_{d,1}(dy, x\gamma_d) \\
& + \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \mu_d(y, x\gamma_d) \right\} \nu_{n,d,1}(dy, x\gamma_d) + \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \mu_d(y, x\gamma_d) \right\} \mu_{d,1}(dy, x\gamma_d).
\end{aligned}$$

Applying arguments similar to those from Lemma 3.1 of Lopez (2011) and Lemma A.2 of Fan and Liu (2018) yields that the last three terms are asymptotically dominated by the first one due to under-smoothing. Hence, we provide detailed derivation for the first term only. Let

$$
\begin{aligned}
L_{n,6} &= \int_0^t \left\{ \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \nu_{n,d}(y,x\gamma_d) \right\} \nu_{n,d,1}(dy, x\gamma_d) \\
&= \frac{1}{n^2h^2} \left\{ \sum_{i=1}^{n} g_4(W_i, W_i, \omega) + \sum_{i \neq j}^{n} g_4(W_i, W_j, \omega) \right\} \equiv L^a_{n,6} + L^b_{n,6},
\end{aligned}
$$

where

$$
\begin{aligned}
g_4(W_1, W_2, \omega) &= \left\{ g_{41}(W_1, \omega) g_{42}(W_2, Y_1, \omega) - \int g_{41}(w_1, \omega) g_{42}(W_2, y_1, \omega) dF(y_1, x_1, d_1, r_1) \right\}, \\
g_{41}(W_1, \omega) &= R_1 \mathbb{1}\{D_1 = d, Y_1 \leq t\} hK_h(x\gamma_d, X_1\gamma_d), \\
g_{42}(W_2, y, \omega) &= \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x) \{ \mathbb{1}\{D_2 = d, Y_2 \leq y \wedge t\} hK_h(x\gamma_d, X_2\gamma_d) \\
&\quad - \int \mathbb{1}\{d_2 = d, y_2 \leq y \wedge t\} hK_h(x\gamma_d, x_2\gamma_d) dF(y_2, x_2, d_2) \bigg\}.
\end{aligned}
$$

Note that the second term is a degenerate second order U process. Lemma 1.7 indicates that the class

$$
\mathscr{G}_4 = \{ (w_1, w_2) \mapsto g_4(w_1, w_2, \omega) : \omega \in \Omega \} \tag{1.9.23}
$$

is of VC type with a bounded envelop. Standard calculation implies that $\sup_{\omega \in \Omega} \left| L^a_{n,6} \right| = O_p(n^{-1}h^{-1})$ and the maximal variance $\sup_{g \in \mathscr{G}_4} \mathbb{E}[g^2]$ is $O\left(h^2\right)$. Another application of Theorem 8 of Giné and Mason (2007) and the Markov inequality yields that $L^b_{n,6}$ is of order $O_p\left(\log n \cdot n^{-1}h^{-1}\right)$ uniformly over Ω.

Gathering results on (1.9.11) - (1.9.14), we conclude that $\sup_{(t,x,\theta) \in \tilde{\mathscr{T}} \times \mathscr{X} \times \Theta} |r_{n,2}(t,x,\theta)| = O_p\left(\log n \cdot n^{-1}h^{-1}\right) = o_p\left(n^{-1/2}\right)$, concluding the proof of Step 2.

To finish the proof, we deduce from a second order Taylor expansion of ϕ_θ^{-1} that, for each $(t,x,\theta) \in \tilde{\mathscr{T}} \times \mathscr{X} \times \Theta$,

$$
\begin{aligned}
\hat{s}_{T_d}(t, x\hat{\gamma}_d, \theta) - s_{T_d}(t, x\gamma_d, \theta) = & \frac{1}{\phi'_\theta(s_{T_d}(t, x\gamma_d, \theta))} \left(\phi_\theta\left(\hat{s}_{T_d}(t, x\hat{\gamma}_d, \theta)\right) - \phi_\theta\left(s_{T_d}(t, x\gamma_d, \theta)\right) \right) \\
& - \frac{\ddot{\phi}_\theta^{-1}(\tilde{s}_d(t,x,\theta))}{\dot{\phi}_\theta^{-1}(\tilde{s}_d(t,x,\theta))^3} \left(\phi_\theta\left(\hat{s}_{T_d}(t, x\hat{\gamma}_d, \theta)\right) - \phi_\theta\left(s_{T_d}(t, x\gamma_d, \theta)\right) \right)^2,
\end{aligned}
$$

where the random function $\tilde{s}_d(t,x,\theta)$ lies between $\hat{s}_{T_d}(t, x\hat{\gamma}_d, \theta)$ and $s_{T_d}(t, x\gamma_d, \theta)$. From Assumption 1.6.4, it holds that both $1/\dot{\phi}'_\theta(z)$ and $\ddot{\phi}_\theta^{-1}(z)$ are uniformly bounded when $z \in [0, y^*_o]$. Also, by the definition of y^*_o, the event $\mathbb{1}\{\tilde{s}_d(t,x,\theta) \leq y^*_o\}$ has the asymptotic probability equal to one, uniformly in (t,x,θ). As a result, the second term is asymptotically negligible. This concludes the proof. ■

1.9.1.2.2 Weak Convergence of CBGP and UBGP

Proof of Corollary 1.1. The uniform representation from Theorem 1.3 consists of four parts. Standard analysis using maximal inequality implies that $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|n^{-1}\sum_{i=1}^{n}\eta_{l,d}(W_i,x,t,\theta)\right| = O_p(n^{-1/2})$. Multiplying the quantity by $\sqrt{nh}$ gives a rate of $O_p\left(h^{1/2}\right) = o_p(1)$ for the second part. Moreover, $\sqrt{nh}\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta} r_n(x,t,\theta) = O_p\left((\log n)^{1/2} n^{-1/2}h^{-1}\right) = o_p(1)$ under Assumption 1.9.2. Next, we show that

$$\sqrt{nh}\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|n^{-1}\sum_{i=1}^{n}\eta_{b,d}(W_i,x,t,\theta)\right| = o_p(1).$$

Let $\tilde{\eta}_{b,d} \equiv \eta_{b,d} - \mathbb{E}[\eta_{b,d}]$ denote the centered version of $\eta_{b,d}$. Standard bias calculation shows that $\mathbb{E}[\eta_{b,d}(W,x,t,\theta)] = O(h^s)$. Under Assumption 1.6, the rate of bias holds uniformly in $(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta$. Define

$$\begin{aligned}\mathscr{G}_b \equiv \{\tilde{w} \mapsto K(x\gamma_d,\tilde{x}\gamma_d)\Psi_d(G_d(\cdot,\tilde{x}\gamma_d) - G_d(\cdot,x\gamma_d), G_{d,1}(\cdot,\tilde{x}\gamma_d) - G_{d,1}(\cdot,x\gamma_d))(t,x,\theta) :\\ (t,\theta)\in\tilde{\mathscr{T}}\times\Theta\}. \qquad (1.9.24)\end{aligned}$$

From Lemma 1.7, this is a VC type class with bounded entropy. We show below that its maximum variance is $O(h^2)$. It suffices to illustrate on the first part, i.e.

$$\begin{aligned}&\mathbb{E}\left[K(x\gamma_d,X\gamma_d)^2\left(\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{G_d(y,X\gamma_d) - G_d(y,x\gamma_d)\right\}s_{d,1}(dy,x\gamma_d)\right)^2\right]\\ &= h^3\int_{\mathbb{R}} u^2k(u)^2du\cdot\left(\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\partial_{x\gamma}G_d(y,\check{x}\gamma_d)s_{d,1}(dy,x\gamma_d)\right)^2\\ &\leq h^3\int_{\mathbb{R}} u^2k(u)^2du\cdot\sup_{(u,\theta)\in[\upsilon_o,1]\times\Theta}\left|\phi^{''}_{\theta}(u)\right|^2\sup_{(y,x)\in\tilde{\mathscr{T}}\times\mathscr{X}}\left\{\left|\partial_{x\gamma}G_d(y,x\gamma_d)\right|^2\left|s_{d,1}(y,x\gamma_d)\right|^2\right\} = O(h^3),\end{aligned}$$

where the first equality is due to the mean value theorem and a change of variable. The inequality follows under Assumption 1.6. It follows by Lemma 1.5 and after multiplying by h^{-1}, that $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|n^{-1}\sum_{i=1}^{n}\eta_{b,d}(W_i,x,t,\theta)\right|$ $= O_p\left((\log n)^{1/2}n^{-1/2}h^{1/2}\right)$. Combining with the result on bias, we have $\sqrt{nh}\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}|n^{-1}\sum_{i=1}^{n}\eta_{b,d}(W_i,$ $x,t,\theta)| = O_p\left((\log n)^{1/2}h + n^{1/2}h^{s+1/2}\right)$, which is $o_p(1)$ under Assumption 1.9.2.

Hence, it remains to prove the weak convergence of

$$\hat{\mathbb{G}}_n^{x\dagger} \equiv \sum_{i=1}^{n} f_{ni}^{x}(t,\theta,d),$$

where $f_{ni}^{x}(t,\theta,d) \equiv n^{-1/2}h^{1/2}\eta_{s,d}(W_i,x,t,\theta)$, for each $(d,t,\theta)\in\{0,1\}\times\tilde{\mathscr{T}}\times\Theta$ and given a fixed value x. This task can be accomplished by invoking the functional central limit theorem for non-identically distributed stochastic process as presented in Lemma 1.6. Note that under Assumption 1.5.1, the triangular array $\{f_{ni}^{x}(t,\theta,d)\}$ is row-wise independent. By definition of $\eta_{s,d}(W_i,x,t,\theta)$, and Assumption 1.6, all of the array components $f_{ni}^{x}(t,\theta,d)$ are right continuous in both t and θ, which implies that the triangular array is separable. Consequently, $\{f_{ni}^{x}(t,\theta,d)\}$ is AMS by Lemma 2 of Kosorok (2003).

To verify manageability, we first note that

$$\mathscr{G}_{\eta} = \{\tilde{w} \mapsto K(x\gamma_d,\tilde{x}\gamma_d)\Psi_d(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d})(t,x,\theta) : (t,\theta) \in \tilde{\mathscr{T}} \times \Theta\}, \tag{1.9.25}$$

is a VC class with an envelop $\sum_{d=0,1} H_{\eta,d}(\tilde{x}\gamma_d)$ by Lemma 1.7, where $0 \leq H_{\eta,d}(\cdot) < M$ for all $\tilde{x} \in \mathscr{X}$ and a positive constant M. Multiplying $n^{-1/2}h^{-1/2}$ preserves the VC property and we conclude by Theorem 11.21 in Kosorok (2008) that $\{f_{ni}\}$ is manageable with the envelop $\{F_{ni}\}$, where $F_{ni}(\tilde{w}) = n^{-1/2}h^{-1/2}\sum_{d=0,1} H_{\eta,d}(\tilde{x}\gamma_d)$, for all $i = 1,...,n$.

For condition (ii), we define $\chi_{n,d}^{x}(t,\theta) = \sum_{i=1}^{n} f_{ni}^{x}(t,\theta,d) - \mathbb{E}[f_{ni}^{x}(t,\theta,d)]$. As a result of the independence of W_i and W_j when $i \neq j$ and the fact that $\mathbb{E}[f_{ni}^{x}(t,\theta,d)] = 0$,

$$\mathbb{E}[\chi_{n,d_1}^{x}(t_1,\theta_1)\chi_{n,d_2}^{x}(t_2,\theta_2)] = \sum_{i=1}^{n} \mathbb{E}\left[f_{ni}^{x}(t_1,\theta_1,d_1)f_{ni}^{x}(t_2,\theta_2,d_2)\right].$$

Furthermore, the right hand side is identically zero if $d_1 \neq d_2$ due to the definition of $\mathscr{E}_{d,\gamma}$ and $\mathscr{E}_{d,1,\gamma}$. Condition (ii) is trivially satisfied in this case, and thus we focus on $d_1 = d_2 = d$. From direct calculations in Section 1.9.3.3, we find

$$\mathbb{E}\left[f_{ni}^{x}(t_1,\theta_1,d)f_{ni}^{x}(t_2,\theta_2,d)\right] = \frac{1}{n}\sigma_{d,x}^{2}(t_1,\theta_1,t_2,\theta_2) + O(n^{-1}h). \tag{1.9.26}$$

where $\sigma_{d,x}^{2}$ is defined in (1.9.38). Under Assumptions 1.8.1 and 1.6, $\|K\|$, $\phi_{\cdot}^{'}$, $\phi_{\cdot}^{''}$, G_d and $G_{d,1}$ are all uniformly bounded. In addition, $f_d(x\gamma_d)$ and $\phi_{\cdot}^{'}(s_{T_d}(\cdot,x\gamma_d,\cdot))$ are uniformly bounded away from zero for each $x \in \mathscr{X}$, under Assumptions 1.3.1 and 1.6.4. Since $h \to 0$ as $n \to \infty$, $\lim_{n\to\infty} \mathbb{E}[\chi_{n,d_1}^{x}(t_1,\theta_1)\chi_{n,d_2}^{x}(t_2,\theta_2)] = \sigma_{d,x}^{2}(t_1,\theta_1,t_2,\theta_2)$ and the limit is well-defined. As a result, condition (ii) holds.

Next, condition (iii) follows from the fact that

$$\sum_{i=1}^{n} \mathbb{E}^{*}[F_{ni}^{2}] \leq 2\sum_{d=0,1}\int_{\mathscr{X}_{\Gamma}} h^{-1}H_{\eta,d}^{2}(\tilde{x}\gamma_d)f(\tilde{x}\gamma_d)d\tilde{x}\gamma_d = 2\sum_{d=0,1} C_d^{2}\int_{[-1,1]} f(x\gamma_d + uh)du < \infty,$$

where the second equality follows from a change of variable, and the last inequality is due to $H_{\eta,d}$ being uniformly bounded.

Regarding the Lindeberg type condition (iv), note that

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}^*[F_{ni}^2\mathbb{1}\{F_{ni}>\varepsilon\}]\\
=&\int_{\mathscr{X}_{\gamma_0}}\int_{\mathscr{X}_{\gamma_1}} h^{-1}\left(\sum_{d=0,1}H_{\eta,d}(\tilde{x}\gamma_d)\right)^2\mathbb{1}\left\{n^{-1/2}h^{-1/2}\sum_{d=0,1}H_{\eta,d}(\tilde{x}\gamma_d)>\varepsilon\right\}f(\tilde{x}\gamma_1,\tilde{x}\gamma_0)d\tilde{x}\gamma_1 d\tilde{x}\gamma_0\\
=&\int_{\mathbb{R}}\int_{\mathbb{R}} h\left(\sum_{d=0,1}C_d\mathbb{1}\{|u_d|\leq 1\}\right)^2\\
&\cdot\mathbb{1}\left\{n^{-1/2}h^{-1/2}\sum_{d=0,1}C_d\mathbb{1}\{|u_d|\leq 1\}>\varepsilon\right\}f(x\gamma_1+u_1h,x\gamma_0+u_0h)du_1du_0.
\end{aligned}$$

Since $nh\to\infty$, the limit of the right hand side as $n\to\infty$ equals zero for each ε by the dominated convergence theorem. Thus condition (iv) is satisfied.

In view of (1.9.26), we obtain from expanding the square in $\rho_n(s,t)$ that $\rho_n(t_1,\theta_1,t_2,\theta_2)=\rho(t_1,\theta_1,t_2,\theta_2)+O(h)$, with $\rho(t_1,\theta_1,t_2,\theta_2)=\left\{\sigma_{d,x}^2(t_1,t_1,\theta_1,\theta_1)-2\sigma_{d,x}^2(t_1,t_2,\theta_1,\theta_2)+\sigma_{d,x}^2(t_2,t_2,\theta_2,\theta_2)\right\}^{1/2}$ for each $(t_1,t_2,\theta_1,\theta_2)\in\tilde{\mathscr{T}}^2\times\Theta^2$. Since the second term vanishes as $n\to\infty$ and the first one is independent of n, we have $\rho(t_{1,n},\theta_{1,n},t_{2,n},\theta_{2,n})\to 0$ implies $\rho_0(t_{1,n},\theta_{1,n},t_{2,n},\theta_{2,n})\to 0$, for all deterministic sequences of $\{t_{1,n},\theta_{1,n}\}$ and $\{t_{2,n},\theta_{2,n}\}$.

We have shown that the triangular array $\{f_{ni}\}$ satisfies conditions (i) - (v) of Lemma 1.6, which implies that $\hat{\mathbb{G}}_n^{x\dagger}$ converges weakly to a two-dimensional Gaussian process with covariance function $\Sigma_\eta^{x\dagger}(\cdot,\cdot)$. Lemma 1.10 shows that $\Sigma_\eta^{x}(\cdot,\cdot)=\Sigma_\eta^{x\dagger}(\cdot,\cdot)+o(1)$. Combining this result with the fact that $\hat{\mathbb{G}}_n^{x}-\hat{\mathbb{G}}_n^{x\dagger}=o_p(1)$ concludes the proof. ■

Proof of Corollary 1.2. Proof of part (i). In view of the uniform representation from Theorem 1.3, we obtain

$$\begin{aligned}
\hat{s}_{T_d}(t,\theta)-s_{T_d}(t,\theta)=&\mathbb{E}_n\left[s_{T_d}(y,X\gamma_d,\theta)-\mathbb{E}[s_{T_d}(y,X\gamma_d,\theta)]\right]\\
&+\mathbb{E}_n\left[\hat{s}_{T,d}(y,X\hat{\gamma}_d,\theta)-s_{T_d}(y,X\gamma_d,\theta)\right]\\
\equiv&A_{n,1}+A_{n,2}.
\end{aligned}$$

The first term is an empirical process indexed by $\varphi_{d,2}$. Utilizing the uniform linear representation in Theorem 1.3, we can further decompose $A_{n,2}$ as

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{h^{-1}g_5(W_i,W_j,t,h,\theta)+g_6(W_i,W_j,\theta)+\eta_{b,d}(W_i,X_j,t,\theta)\right\}+\frac{1}{n}\sum_{i=1}^{n}+r_n(X_i,t,\theta),$$

where

$$
\begin{aligned}
g_5(W_1,W_2,t,h,\theta) &= \frac{hK_h(X_2\gamma_d,X_1\gamma_d)}{f_d(X_2\gamma_d)}\Psi(\mathscr{E}_{d,\gamma_d,1},\mathscr{E}_{d,1,\gamma_d,1})(t,X_2,\theta),\\
g_6(W_1,W_2,\theta) &= \frac{\psi_d^b(X_1)'V_d^{-1}}{f_d(X_2\gamma_d)}\Psi(\psi_d^a,\psi_{d,1}^a)(t,X_2,\theta).
\end{aligned}
$$

From the proof of Corollary 1.1, we find that $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|n^{-1}\sum_{i=1}^n\eta_{b,d}(W_i,x,t,\theta)\right| = O_p\left((\log n)^{1/2}n^{-1/2}h^{1/2}\right)$ $+O(h^s)$, thus the double mean involving $\eta_{b,d}$ is uniformly $o_p\left(n^{-1/2}\right)$ under Assumption 1.9.2. Additionally, from Theorem 1.3 we have $r_n(x,t,\theta) = o_p\left(n^{-1/2}\right)$ uniformly over $\mathscr{X}\times\tilde{\mathscr{T}}\times\Theta$, thus the last term is also asymptotically negligible.

Consequently, it suffices to work on the first two terms. We will show (a) the U-process indexed by g_5 is asymptotically equivalent to an empirical process indexed by $\mathbb{E}[g_5|W_1]$, and (b) the U-process indexed by $\eta_{l,d}$ is asymptotically negligible.

We focus on (a) first. It is straightforward to show that $\Psi(\mathscr{E}_{d,\gamma_d,1},\mathscr{E}_{d,1,\gamma_d,1})/f_d(x\gamma_d)$ is uniformly bounded. We therefore deduce from direct calculations that $\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|\frac{1}{n^2h}\sum_{i=1}^n g_5(W_i,W_i,t,h,\theta)\right| = O_p\left(n^{-1}h^{-1}\right)$. Observe that, by Lemma 1.7,

$$
\mathscr{G}_5 = \left\{(w_1,w_2)\mapsto g_5(w_1,w_2,t,h,\theta) : (t,h,\theta)\in\tilde{\mathscr{T}}\times\mathscr{H}\times\Theta\right\}, \tag{1.9.27}
$$

is of VC type with bounded envelop. Also, $\mathbb{E}[g_5|W_2] = 0$ and $\sup_{g\in\mathscr{G}_5}\mathbb{E}[g^2] = O(h)$. As a result, we may deduce from the maximal inequality in Lemma 1.5 that

$$
\begin{aligned}
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}&\left|\left\{\frac{1}{n(n-1)h}\sum_{i\neq j}^n g_5(W_i,W_j,t,h,\theta) - \frac{1}{nh}\sum_{i=1}^n\mathbb{E}[g_5(W_i,W_j,t,h,\theta)|W_i]\right\}\right|\\
&= O_p\left(\log n\cdot n^{-1}h^{-1/2}\right) = o_p\left(n^{-1/2}\right),
\end{aligned}
$$

which implies that the second order U process can be uniformly approximated by an empirical process indexed by the conditional mean. Let $A_{n,21}$ denote this process, for which we have the following

$$
\begin{aligned}
&\mathbb{E}[g_5(W_1,W_2,t,h,\theta)|W_1]\\
&=h\int\frac{K_h(x\gamma_d,X_1\gamma_d)}{f_d(x\gamma_d)}\Psi(\mathscr{E}_{d,\gamma_d,1},\mathscr{E}_{d,1,\gamma_d,1})(t,x,\theta)f(x\gamma_d)d(x\gamma_d)\\
&=h\int\frac{K(u)}{f_d(X_1\gamma_d+uh)\phi_\theta'\left(s_{T_d}(t,X_1\gamma_d+uh,\theta)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\cdot\left\{\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y, X_1\gamma_d+uh)\mathbb{1}\{D_1=d\}\left(\mathbb{1}\{Y_1\leq y\}-G_d(y,X_1\gamma_d)\right)s_{d,1}(dy,X_1\gamma_d+uh)\right.\\
&-\phi'_{\theta}\left(s_d(t,X_1\gamma_d+uh)\right)\mathbb{1}\{D_1=d\}\left(R_1\mathbb{1}\{Y_1\leq t\}-G_{d,1}(t,X_1\gamma_d)\right)\\
&-\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,X_1\gamma_d+uh)\mathbb{1}\{D_1=d\}\left(R_1\mathbb{1}\{Y_1\leq y\}-G_{d,1}(y,X_1\gamma_d)\right)\\
&\left.\cdot s_d(dy,X_1\gamma_d+uh)f(X_1\gamma_d+uh)\right\}du\\
=&\frac{hf(X_1\gamma_d)}{f_d(X_1\gamma_d)}\Psi(\mathscr{E}_{d,\gamma_d,1},\mathscr{E}_{d,1,\gamma_d,1})(t,X_1,\theta)+O\left(h^{s+1}\right).
\end{aligned}
$$

The second equality follows by a change of variable and the last one is due to Assumption 1.8.1. Since $\phi^{'}_{\theta}(\cdot), \phi^{''}_{\theta}(\cdot)$, $G_d(y,\cdot), G_{d,1}(y,\cdot)$, $f_d(\cdot)$, and $f(\cdot)$ are all $(s+1)$ times continuously differentiable with uniformly bounded derivatives under Assumption 1.6, the rate of the bias holds uniformly over $\tilde{\mathscr{T}}\times\Theta$.

Now, we show (b). Observe that g_6 is multiplicatively separable in W_1 and W_2. The part involving W_1 is $O_p\left(n^{-1/2}\right)$ and not indexed by (t,θ), it then suffices to show that the empirical process indexed by $g_{61}(W_2,t,\theta)\equiv \Psi_d(\psi^a_d,\psi^a_{d,1})(t,X_2,\theta)/f_d\left(X_2\gamma_d\right)$ is $o_p(1)$. Note that

$$
\begin{aligned}
\mathbb{E}[\psi^a_d(y,X_2)|X_2\gamma_d] &= \mathbb{E}[\partial_{x\gamma}G_d(y,X_2\gamma_d)(\mathbb{E}_{X_1\gamma_d}[X_1|X_2\gamma_d]-X_2)|X_2\gamma_d]\\
&= \partial_{x\gamma}G_d(y,X_2\gamma_d)(\mathbb{E}_{X_1\gamma_d}[X_1|X_2\gamma_d]-\mathbb{E}_{X_2\gamma_d}[X_2|X_2\gamma_d])=0,
\end{aligned}
$$

where the last equality holds because X_1 and X_2 are identically distributed. The same result holds for $\psi^a_{d,1}$. By Fubini's theorem, and the law of iterated expectation, it follows that $\mathbb{E}[g_{61}]=0$. Next, let

$$
\mathscr{G}_6=\left\{(w_1,w_2)\mapsto g_{61}(w_1,w_2,t,\theta):(t,\theta)\in\tilde{\mathscr{T}}\times\Theta\right\}. \tag{1.9.28}
$$

Lemma 1.7 establishes that it is of the VC type with bounded envelop. Moreover, its maximal variance is $O(1)$. We deduce from Lemma 1.5 that $\mathbb{E}_n[g_{61}(W_2,t,\theta)]=O_p\left(n^{-1/2}\right)$ uniformly over $\tilde{\mathscr{T}}\times\Theta$. Overall, the U process indexed by g_6 is $O_p(n^{-1})$, and thus, asymptotically negligible.

Proof of part (ii). In view of the uniform representation established in the previous part, weak convergence follows from Theorem 2.1 of Kosorok (2008) if the class of function

$$
\mathscr{G}_{\varphi}\equiv\{w\mapsto\varphi_d(w,t,\theta):(d,t,\theta)\in\{0,1\}\times\tilde{\mathscr{T}}\times\Theta\} \tag{1.9.29}
$$

is Donsker. We see from Lemma 1.7 that $\mathscr{G}_{\varphi}$ of VC type, which implies that it is Donsker by Theorem 19.14 in Van der Vaart (1998). ■

1.9.1.2.3 Functional Delta Method

Before proving Theorem 1.4, we state a general result on functional delta method. Let $\nu(\cdot)$ denote a generic functional mapping from $\ell_\infty(\tilde{\mathscr{T}} \times \Theta^2) \times \ell_\infty(\tilde{\mathscr{T}} \times \Theta^2)$ to a normed space $\ell_\infty(\mathscr{U} \times \Theta^2) \times \ell_\infty(\mathscr{U} \times \Theta^2)$.

Lemma 1.2 (i) Suppose this functional of interest is Hadamard differentiable at $\mathbf{S}^x$, for a fix $x \in \mathscr{X}$, tangentially to a space $\mathscr{C}(\mathscr{U} \times \Theta^2)$ with derivative $\nu'_{\mathbf{S}^x}$, and that the assumptions of Corollary 1.1 hold, then

$$\sqrt{nh}\left(\nu\left(\hat{\mathbf{S}}^x\right)(\cdot,\cdot) - \nu\left(\mathbf{S}^x\right)(\cdot,\cdot)\right) \Rightarrow \nu'_{\mathbf{S}^x}(\mathbb{G})(\cdot,\cdot) \equiv \mathbb{G}^x_\nu,$$

in $\ell_\infty(\mathscr{U} \times \Theta^2) \times \ell_\infty(\mathscr{U} \times \Theta^2)$, where $\mathbb{G}^x_\nu$ is a two-dimensional Gaussian process with zero mean and covariance function,

$$\Sigma^x_\nu(\mathbf{u},\boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{\varphi}^x_\nu(W,u_1,\boldsymbol{\theta}_1)\boldsymbol{\varphi}^x_\nu(W,u_2,\boldsymbol{\theta}_2)'\right],$$

where $\boldsymbol{\varphi}^x_\nu \equiv \nu'_{\mathbf{S}^x}(\boldsymbol{\varphi}^x)$, for each $\mathbf{u} = (u_1,u_2)' \in \mathscr{U} \times \mathscr{U}$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)' \in \Theta^2 \times \Theta^2$.

(ii) Suppose $\nu(\cdot)$ is Hadamard differentiable at $\mathbf{S}$ tangentially to a space $\mathscr{C}(\mathscr{U} \times \Theta^2)$ with derivative $\nu'_{\mathbf{S}}$, and that the assumptions of Corollary 1.2 hold, then

$$\sqrt{n}\left(\nu\left(\hat{\mathbf{S}}\right)(\cdot,\cdot) - \nu\left(\mathbf{S}\right)(\cdot,\cdot)\right) \Rightarrow \nu'_{\mathbf{S}}(\mathbb{G})(\cdot,\cdot) \equiv \mathbb{G}_\nu,$$

in $\ell_\infty(\mathscr{U} \times \Theta^2) \times \ell_\infty(\mathscr{U} \times \Theta^2)$, where $\mathbb{G}_\nu$ is a two-dimensional Gaussian process with zero mean and covariance function,

$$\Sigma_\nu(\mathbf{u},\boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{\varphi}_\nu(W,u_1,\boldsymbol{\theta}_1)\boldsymbol{\varphi}_\nu(W,u_2,\boldsymbol{\theta}_2)'\right],$$

where $\boldsymbol{\varphi}_\nu \equiv \nu'_{\mathbf{S}}(\boldsymbol{\varphi})$, for each $\mathbf{u} = (u_1,u_2)' \in \mathscr{U} \times \mathscr{U}$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)' \in \Theta^2 \times \Theta^2$.

Proof of Theorem 1.4. We establish Hadamard differentiability for each type of treatment effects and the desired result would then follow by a direct application of Lemma 1.2. The cases for DTE and CHTE follow immediately from Lemma 3.9.25 in Van Der Vaart and Wellner (1996). Next, note that integration with respect to the Lebesgue measure is a linear operator, which implies that $\boldsymbol{\nu}_{ATE}(\cdot,\cdot)$ is linear in (s_{T_1}, s_{T_0}), and thus, is Hadamard differentiable by definition.

For QTE, it suffices to show that the mapping $q_{d,\cdot}(\cdot) : \ell_\infty(\tilde{\mathscr{T}} \times \Theta) \mapsto \ell_\infty((0,\bar{\tau}) \times \Theta)$ is Hadamard differentiable. The proof is similar to that of Lemma 3.9.23 in Van Der Vaart and Wellner (1996). Let $h_t \to h$ uniformly in $\ell_\infty(\tilde{\mathscr{T}} \times \Theta)$, with h being continuous. Thus, $s_{T_d} + th_t \in \ell_\infty(\tilde{\mathscr{T}} \times \Theta)$ for all $t > 0$. Let $q_{d,\theta,t}(\tau) \equiv \inf\{y : (s_{T_d} + th_t)(y,\theta) \leq 1 - \tau\}$. Due to $s_{T_d}(\cdot,\theta)$ and $(s_{T_d} + th_t)(\cdot,\theta)$ being restricted to $\tilde{\mathscr{T}}$ for each θ, it holds that $q_{d,\theta,t}(\tau), q_{d,\theta,t}(\tau) \in \tilde{\mathscr{T}}$, for all

$(\tau,\theta)\in(0,\tau_o)\times\Theta$. By the definition of $q_{d,\theta,t}$, we have

$$1-(s_{T_d}+th_t)(q_{d,\theta,t}(\tau)-\varepsilon_{d,t,\theta}(\tau),\theta)\le\tau\le 1-(s_{T_d}+th_t)(q_{d,\theta,t}(\tau),\theta),$$

where $\varepsilon_{d,t,\theta}(\tau)=t^2\wedge q_{d,\theta,t}(\tau)>0$. Under Assumptions 1.6.2 and 1.6.4, $s_{T_d}(q_{d,\theta,t}(\tau)-\varepsilon_{d,t,\theta}(\tau),\theta)=s_{T_d}(q_{d,\theta,t}(\tau),\theta)$ $+O(\varepsilon_{d,t,\theta}(\tau))$, uniformly in $(\tau,\theta)\in(0,\tau_o)\times\Theta$. This further implies that

$$-th(q_{d,\theta,t}(\tau)-\varepsilon_{d,t,\theta}(\tau))+r_t(\tau,\theta)\le s_{T_d}(q_{d,\theta,t}(\tau),\theta)-s_{T_d}(q_{d,\theta}(\tau),\theta)\le -th(q_{d,\theta,t}(\tau))+r_t(\tau,\theta),$$

where $r_t(\tau,\theta)=o(t)$, uniformly in τ and θ. From the continuous differentiability of $s_{T_d}(y,\theta)$ in y and the derivative $-f_{T_d}(y,\theta)$ being uniformly bounded, we deduce that $\left|q_{d,\theta,t}(\tau)-q_{d,\theta}(\tau)\right|=O(t)$, uniformly in τ and θ. Applying Taylor expansion of the middle term in the preceding display allows us to conclude that $q_{d,\cdot}(\cdot)$ is Hadamard differentiable at $s_{T_d}(\cdot,\cdot)$, with derivative given by $h\mapsto h(q_{d,\cdot}(\cdot),\cdot)/f_{T_d}(q_{d,\cdot}(\cdot),\cdot)$, for $d\in\{0,1\}$. Another application of Lemma 3.9.25 in Van Der Vaart and Wellner (1996) yields the Hadamard differentiability of $\boldsymbol{\nu}_{QTE}(\cdot,\cdot)$, concluding our proof.

■

1.9.1.3 Proofs for Results from Section 1.5

1.9.1.3.1 Weak Convergence of Multiplier Bootstrap Processes

Proof of Theorem 1.5. Proof of part (i) We first show that $\mathbb{G}_{n,\xi}\overset{p}{\underset{\xi}{\rightsquigarrow}}\mathbb{G}$. In view of Theorem 11.19 in Kosorok (2008), the conditional weak convergence follows under Assumption 1.11 if the triangular array $\{f^x_{ni}\}_{i=1}^n$, with $f^x_{ni}(t,\theta,d)=n^{-1/2}h^{1/2}\eta_{s,d}(W_i,x,t,\theta)$, satisfies the conditions of Lemma 1.6. Note that we have verified these conditions in Corollary 1.1. Hence, the desired result holds. Next, we prove

$$\sup_{(t,\boldsymbol{\theta})\in\tilde{\mathscr{T}}\times\Theta^2}\left\|\hat{\mathbb{G}}^x_{\xi}(t,\boldsymbol{\theta})-\mathbb{G}^x_{n,\xi}(t,\boldsymbol{\theta})\right\|=o_p(1). \tag{1.9.30}$$

Decompose $\hat{\eta}_{s,d}(W,\xi,x,t,\theta)$ as

$$\begin{aligned}&\frac{K_h(x\hat{\gamma}_d,X\hat{\gamma}_d)-K_h(x\gamma_d,X\gamma_d)}{f_d(x\gamma_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\\ &-\frac{\left(\hat{f}_d(x\hat{\gamma}_d)-f_d(x\gamma_d)\right)K_h(x\gamma_d,X\gamma_d)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\\ &+\frac{K_h(x\gamma_d,X\gamma_d)}{f_d(x\gamma_d)}\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,x,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\right\}\end{aligned}$$

$$
\begin{aligned}
&-\frac{\left(\hat{f}_d(x\hat{\gamma}_d)-f_d(x\gamma_d)\right)\left(K_h(x\hat{\gamma}_d,X\hat{\gamma}_d)-K_h(x\gamma_d,X\gamma_d)\right)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\\
&-\frac{\left(\hat{f}_d(x\hat{\gamma}_d)-f_d(x\gamma_d)\right)K_h(x\gamma_d,X\gamma_d)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,x,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\right\}\\
&+\frac{K_h(x\hat{\gamma}_d,X\hat{\gamma}_d)-K_h(x\gamma_d,X\gamma_d)}{f_d(x\gamma_d)}\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,x,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\right\}\\
&-\frac{\left(\hat{f}_d(x\hat{\gamma}_d)-f_d(x\gamma_d)\right)\left(K_h(x\hat{\gamma}_d,X\hat{\gamma}_d)-K_h(x\gamma_d,X\gamma_d)\right)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\\
&\quad\cdot\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,x,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\right\}\\
=&A_{n,1}(W,\xi)+A_{n,2}(W,\xi)+A_{n,3}(W,\xi)+A_{n,4}(W,\xi)+A_{n,5}(W,\xi)+A_{n,6}(W,\xi)+A_{n,7}(W,\xi).
\end{aligned}
$$

The first three terms are obtained from the "first-order" expansion, the rate of which is dominating. Crude bounds based on uniform rates from Lemma 1.4 can be utilized to control the remaining three terms.

By a Taylor expansion, we have

$$
\begin{aligned}
A_{n,1}(W,\xi)=&\frac{h^{-1}K_h^{(1)}(x'\gamma_d,X_i'\gamma_d)(x_{[-1]}-X_{[-1],i})'(\hat{\gamma}_d-\gamma_d)}{f_d(x\gamma_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\\
&+\frac{h^{-2}K_h^{(2)}(x'\tilde{\gamma}_d,X_i'\tilde{\gamma}_d)((x_{[-1]}-X_{[-1],i})'(\hat{\gamma}_d-\gamma_d))^2}{f_d(x\gamma_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,x,\theta)\\
=&A_{n,11}(W,\xi)+A_{n,12}(W,\xi).
\end{aligned}
$$

Due to the independence of the bootstrap weights, $\mathbb{E}[A_{n,1}(X,\xi)|X]=0$, the empirical process $n^{-1/2}h^{1/2}\sum_{j=1}^{n}A_{n,11}(W_j,\xi_j)$ is centered, and by Lemma 1.5, is of the order $O_p\left(\log n\cdot n^{-1/2}h^{-1}\right)=o_p(1)$, uniformly in t and θ. For the second term, using the uniform boundedness of $K^{(2)}$, $\mathscr{E}_{d,\gamma}$, and the compactness of $\mathscr{X}$, we deduce that

$$
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}h^{1/2}\sum_{j=1}^{n}A_{n,11}(W_j,\xi_j)\right|=o_p(n^{-1/2}h^{1/2}).
$$

Note that $n^{-1/2}h^{1/2}\sum_{j=1}^{n}A_{n,2}(W_j,\xi_j)$ can be bounded by

$$
\begin{aligned}
&n^{1/2}h^{1/2}\left|\frac{\hat{f}_d(x\hat{\gamma}_d)-f_d(x\gamma_d)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}K_h(x\gamma_d,X_i\gamma_d)\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d,i},\mathscr{E}_{d,1,\gamma_d,i}\right)(t,x,\theta)\right|\\
&=n^{1/2}h^{1/2}O_p\left((\log n)^{1/2}n^{-1/2}h^{-1/2}\right)\cdot O_p\left((\log n)^{1/2}n^{-1/2}h^{-1/2}\right)=O_p\left(\log n\cdot n^{-1/2}h^{-1/2}\right),
\end{aligned}
$$

which is $o_p(1)$ uniformly in t and θ.

From Lemma 1.9 with $\ell=0$, we deduce that $\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}h^{1/2}\sum_{j=1}^{n}A_{n,3}(W_j,\xi_j)\right|=O_p\left(\log n\cdot n^{-1/2}h^{-1/2}\right)$.

Next, for $s_1 = 4,5$, $n^{-1/2}h^{1/2}\sum_{j=1}^{n} A_{n,s_1}(W_j,\xi_j)$ are bounded by

$$\left|\frac{\hat{f}_d(x\hat{\gamma}_d) - f_d(x\gamma_d)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\right| \cdot \left|n^{-1/2}h^{1/2}\sum_{j=1}^{n} A_{n,s_2}(W_j,\xi_j)\right|,$$

for $s_2 = 1,3$, respectively, both of which converge uniformly at a rate of $O_p\left(\log n \cdot n^{-1}h^{-1}\right)$.

By a Taylor expansion of $K_h(x\hat{\gamma}_d, X\hat{\gamma}_d)$ around γ_d, and by applying Lemma 1.9 with $\ell = 1$, we get

$$\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}h^{1/2}\sum_{j=1}^{n} A_{n,6}(W_j,\xi_j)\right| = \|\hat{\gamma}_d - \gamma_d\| \cdot O_p\left((\log n)^{1/2} n^{-1/2}h^{-3/2}\right) = O_p\left((\log n)^{1/2} n^{-1}h^{-3/2}\right).$$

Lastly, $n^{-1/2}h^{1/2}\sum_{j=1}^{n} A_{n,7}(W_j,\xi_j)$ is bounded by

$$\left|\frac{\hat{f}_d(x\hat{\gamma}_d) - f_d(x\gamma_d)}{f_d(x\gamma_d)\hat{f}_d(x\hat{\gamma}_d)}\right| \cdot \left|n^{-1/2}h^{1/2}\sum_{j=1}^{n} A_{n,6}(W_j,\xi_j)\right|,$$

which converges at a rate of $O_p\left(\log n \cdot n^{-3/2}h^{-2}\right)$ uniformly over $\tilde{\mathscr{T}} \times \Theta$. Collecting the results on $A_{n,1}$ to $A_{n,7}$ implies that (1.9.30) holds.

To finish the proof, note that by the triangular inequality,

$$\left|\mathbb{E}_{\xi|w}[h(\hat{\mathbb{G}}^x_{n,\xi})] - \mathbb{E}[h(\mathbb{G}^x)]\right| \leq \left|\mathbb{E}_{\xi|w}[h(\hat{\mathbb{G}}^x_{n,\xi})] - \mathbb{E}_{\xi|w}[h(\mathbb{G}^x_{n,\xi})]\right| + \left|\mathbb{E}_{\xi|w}[h(\mathbb{G}^x_{n,\xi})] - \mathbb{E}[h(\mathbb{G}^x)]\right|,$$

for each $h \in BL_1$. The second term on the right hand side converges to zero since $\mathbb{G}^x_{n,\xi} \underset{\xi}{\overset{p}{\to}} \mathbb{G}^x$. By Jensen's inequality and the definition of BL_1, $\mathbb{E}_{\xi|w}[|h(\hat{\mathbb{G}}^x_{n,\xi}) - h(\mathbb{G}^x_{n,\xi})|] \leq 2\mathbb{P}_{\xi|w}\left(|\hat{\mathbb{G}}^x_{n,\xi} - \mathbb{G}^x_{n,\xi}| > \varepsilon\right) + \varepsilon$, for any $\varepsilon \in (0,1)$. Due to the dominated convergence theorem, the first term on the right hand side goes to zero since $\hat{\mathbb{G}}^x_{n,\xi} - \mathbb{G}^x_{n,\xi} = o_p(1)$. Taking "sup" over BL_1 shows $\hat{\mathbb{G}}^x_{n,\xi} \underset{\xi}{\overset{p}{\to}} \mathbb{G}^x$.

Proof of part (ii) The proof is similar in structure to that of the first part. Thus, we provide a sketch of proof only. In view of Theorem 10.4 in Kosorok (2008), $\mathbb{G}_{n,\xi} \underset{\xi}{\overset{p}{\to}} \mathbb{G}$ provided $\mathscr{G}_\varphi$ as defined in (1.9.29) is a Donsker class. The latter condition is proved in Lemma 1.7. In view of the discussion at the end of the last part, it remains to show that $\sup_{(t,\boldsymbol{\theta})\in\tilde{\mathscr{T}}\times\Theta^2}\left\|\hat{\mathbb{G}}_\xi(t,\boldsymbol{\theta}) - \mathbb{G}_{n,\xi}(t,\boldsymbol{\theta})\right\| = o_p(1)$.

We establish uniform convergence of $\xi \cdot \hat{\varphi}_{d,1}(W,t,\theta)$ first. Decompose the term as

$$\frac{\hat{f}(X\hat{\gamma}_d) - f(X\gamma_d)}{f_d(X\gamma_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d}, \mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)$$

$$
\begin{aligned}
&-\frac{f(X\gamma_d)\left(\hat{f}_d(X\hat{\gamma}_d)-f(X\gamma_d,d)\right)}{f_d(X\gamma_d)\hat{f}_d(X\hat{\gamma}_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)\\
&+\frac{f(X\gamma_d)}{f_d(X\gamma_d)}\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,X,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)\right\}\\
&-\frac{\left(\hat{f}(X\hat{\gamma}_d)-f(X\gamma_d)\right)\left(\hat{f}_d(X\hat{\gamma}_d)-f(X\gamma_d,d)\right)}{f_d(X\gamma_d)^2\hat{f}_d(X\hat{\gamma}_d)}\xi\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)\\
&+\frac{\hat{f}(X\hat{\gamma}_d)-f(X\gamma_d)}{f_d(X\gamma_d)}\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,X,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)\right\}\\
&-\frac{f(X\gamma_d,d)\left(\hat{f}_d(X\hat{\gamma}_d)-f(X\gamma_d,d)\right)}{f_d(X\gamma_d)\hat{f}_d(X\hat{\gamma}_d)}\xi\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,X,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)\right\}\\
&-\frac{\left(\hat{f}_d(X\hat{\gamma}_d)-f(X\gamma_d,d)\right)\left(\hat{f}_d(X\hat{\gamma}_d)-f(X\gamma_d,d)\right)}{f_d(X\gamma_d)\hat{f}_d(X\hat{\gamma}_d)}\xi\\
&\qquad\cdot\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d},\hat{\mathscr{E}}_{d,1,\hat{\gamma}_d}\right)(t,X,\theta)-\Psi_d\left(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}\right)(t,X,\theta)\right\}\\
&=A_{n,1}(W,\xi)+A_{n,2}(W,\xi)+A_{n,3}(W,\xi)+A_{n,4}(W,\xi)+A_{n,5}(W,\xi)+A_{n,6}(W,\xi)+A_{n,7}(W,\xi).
\end{aligned}
$$

It can be shown, via direct analysis in the case of $A_{n,1}$ and $A_{n,2}$ or by further decomposition à la Lemma 1.9 for $A_{n,3}$, that the first three terms are dominated by second order degenerate U processes. Therefore, we can deduce from Lemma 1.5 that

$$
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}\sum_{j=1}^{n}A_{n,\ell}(W_j,\xi_j)\right|=O_p\left(\log n\cdot n^{-1/2}h^{-1/2}\right),
$$

for $\ell=1,2,3$. Similar analysis implies that the $A_{n,4}$-$A_{n,6}$ and $A_{n,7}$ are governed by third order degenerate U processes and a fourth order degenerate U processes, respectively, with

$$
\begin{aligned}
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}\sum_{j=1}^{n}A_{n,\ell}(W_j,\xi_j)\right|&=O_p\left((\log n)^{3/2}n^{-1}h^{-1}\right),\\
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}\sum_{j=1}^{n}A_{n,7}(W_j,\xi_j)\right|&=O_p\left((\log n)^{2}n^{-3/2}h^{-3/2}\right),
\end{aligned}
$$

for $\ell=4,5,6$. We therefore conclude that

$$
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}\sum_{j=1}^{n}\left\{\hat{\varphi}_{d,\xi,1}(W_j,\xi_j,t,\theta)-\varphi_{d,\xi,1}(W_j,\xi_j,t,\theta)\right\}\right|=O_p\left(\log n\cdot n^{-1/2}h^{-1/2}\right)=o_p(1),
$$

for $d=0,1$. Regarding $\xi\hat{\varphi}_{d,2}$, we note that

$$
n^{-1/2}\sum_{j=1}^{n}\left\{\hat{\varphi}_{d,\xi,2}(W_j,\xi_j,t,\theta)-\varphi_{d,\xi,2}(W_j,\xi_j,t,\theta)\right\}
$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{j=1}^{n}\xi_j\left\{\hat{s}_{T,d}(t,X_j\hat{\gamma}_d,\theta)-s_{T_d}(t,X_j\gamma_d,\theta)\right\}\\
&\quad+n^{1/2}\,\mathbb{E}_n[\xi]\cdot\left\{\mathbb{E}_n[\hat{s}_{T,d}(t,X\hat{\gamma}_d,\theta)]-\mathbb{E}[s_{T_d}(t,X\gamma_d,\theta)]\right\}.
\end{aligned}
$$

The term in the second line can be analyzed as in the previous part. Due to Assumption 1.11, we have that $n^{1/2}\,\mathbb{E}_n[\xi]=O_p(1)$. The uniform convergence of $\hat{s}_{T,d}(t,X\hat{\gamma}_d,\theta)$ to $s_{T_d}(t,X\gamma_d,\theta)$ as proved in Theorem 1.3, implies that $\mathbb{E}_n[\hat{s}_{T,d}(t,X\hat{\gamma}_d,\theta)]-\mathbb{E}[s_{T_d}(t,X\gamma_d,\theta)]=o_p(1)$, uniformly over $\tilde{\mathscr{T}}\times\Theta$. Combining the two results, we deduce that the last term in the previous display is also $o_p(1)$, concluding the proof. ■

Proof of Corollary 1.3. Since integration with respect to the Lebesgue measure is a linear, and thus, Lipschitz continuous, mapping, the results for the ATE then follow from Theorem 1.5 and the continuous mapping theorem for multiplier bootstrap, cf. Proposition 10.7 in Kosorok (2008). The case for the DTE also follows trivially from the continuous mapping theorem. Now, let us consider the QTE. Let $\Psi^x_{n,d}(t,\theta)\equiv n^{-1/2}h^{1/2}\sum_{i=1}^{n}\xi_i\cdot\hat{\eta}_{s,d}(W_i,x,t,\theta)$. Using this quantity, we write

$$
\begin{aligned}
&\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|n^{-1/2}h^{1/2}\sum_{i=1}^{n}\xi\left(\hat{\varphi}^x_{d,QTE}(W,\tau,\theta)-\frac{\hat{\eta}_{s,d}(W,x,q^x_{d,\theta}(\tau),\theta)}{f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)}\right)\right| \qquad (1.9.31)\\
=&\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\frac{\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)}{\hat{f}_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)}-\frac{\Psi^x_{n,d}(q^x_{d,\theta}(\tau),\theta)}{f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)}\right|\\
\leq&\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\frac{\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)}{\hat{f}_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)}-\frac{\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)}{f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)}\right|\\
&+\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\frac{\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)}{f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)}-\frac{\Psi^x_{n,d}(q^x_{d,\theta}(\tau),\theta)}{f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)}\right|\\
\lesssim&\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)\right|\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\hat{f}_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)-f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)\right|\\
&+\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)-\Psi^x_{n,d}(q^x_{d,\theta}(\tau),\theta)\right|.
\end{aligned}
$$

From the fact that $\hat{q}^x_{d,\theta}$ converges uniformly to $q^x_{d,\theta}$, and the definition of τ_o, we deduce that $\sup_{\tau\in(0,\tau_o)}\hat{q}^x_{d,\theta}(\tau)\leq y_o$ with probability approaching one. Note that,

$$
\begin{aligned}
&\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\hat{f}_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)-f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)\right|\\
\leq&\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\hat{f}_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)-f_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)\right|+\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|f_{T_d,x}(\hat{q}^x_{d,\theta}(\tau),\theta)-f_{T_d,x}(q^x_{d,\theta}(\tau),\theta)\right|\\
\leq&\sup_{y\in\tilde{\mathscr{T}}}\left|\hat{f}_{T_d,x}(y,\theta)-f_{T_d,x}(y,\theta)\right|+M_1\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\hat{q}^x_{d,\theta}(\tau)-q^x_{d,\theta}(\tau)\right|=o_p(1).
\end{aligned}
$$

Regarding the second inequality, the first term follows by the observation above the display and Assumption 1.12, while the second term follows by the continuous differentiability of $f_{T_d,x}$, which implied by Assumption 1.6.2 and Theorem 1.2. The last equality is due to Assumption 1.12 and the fact that $\hat{q}^x_{d,\theta}(\tau)$ converges to $q^x_{d,\theta}(\tau)$ uniformly over $(0,\tau_o)\times\Theta$.

We have shown, in Theorem 1.5, that $\Psi^x_{n,d}(\cdot,\cdot)$ converges weakly to a centered Gaussian process. It follows immediately, by Prokhorov's theorem and the fact that $\sup_{\tau\in(0,\tau_o)}\hat{q}^x_{d,\theta}(\tau)\leq y_o$ with probability approaching one, that $\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)\right| = O_p(1)$. Moreover, since $\Psi^x_{n,d}(\cdot,\cdot)$ is equicontinuous and that $\hat{q}^x_{d,\theta}$ converges uniformly to $q^x_{d,\theta}$, we get that, conditional on the sample path $\{W_i\}^n_{i=1}$,

$$\sup_{(\tau,\theta)\in(0,\tau_o)\times\Theta}\left|\Psi^x_{n,d}(\hat{q}^x_{d,\theta}(\tau),\theta)-\Psi^x_{n,d}(q^x_{d,\theta}(\tau),\theta)\right| = o_p(1).$$

Collecting the results, we deduce that (1.9.31) $\underset{\xi}{\overset{p}{\to}}$ 0. In addition, by same lines of reasoning as in the proof for Theorem 1.5(i), it is straightforward to show that $\Psi^x_{n,d}(q^x_{d,\cdot}(\cdot),\cdot)/f_{T_d,x}(q^x_{d,\cdot}(\cdot),\cdot)\underset{\xi}{\overset{p}{\to}}\nu'_{QTE,\boldsymbol{S}^x}(\mathbb{G}^x)(\cdot,\cdot)$. This completes the proof for $\hat{\mathbb{G}}^x_{\xi,QTE}$.

To conclude, we note that, the proof for the unconditional QTE and for the DTE's will follow by largely parallel analyses, and thus, we omit it. ■

1.9.1.3.2 Bootstrap Confidence Sets

In this section, we first describe an algorithm for constructing uniform confidence sets for conditional TEBFs. Then, we validate the resulting confidence sets by proving Theorem 1.6. Let $\hat{\mathbb{G}}^x_{lb,\xi,j}$ and $\hat{\mathbb{G}}^x_{ub,\xi,j}$ denote the first and second component of $\hat{\mathbb{G}}^x_{\xi,j}$, respectively.

Algorithm 1.9.1 1. Same as Step 1 of Algorithm 1.9.1. In Steps 2-5, the calculations will be performed for $d,r\in\{0,1\}$, $t\in\tilde{\mathscr{T}}$, $\tau\in(0,\tau_o)$, $\boldsymbol{\theta}\in\boldsymbol{\Theta}_l$, and $u\in\mathscr{U}_m$.

2. Estimate $\hat{\gamma}_d$, $\hat{G}_{d,1}(t,x\hat{\gamma}_d)$, $\hat{G}_d(t,x\hat{\gamma}_d)$, $\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta)$, following (1.4.2), (1.4.3), (1.4.4), and (1.4.5). If $j=QTE$, compute $\hat{q}^x_{d,\theta}(\tau)$ and $\hat{f}_{T_d,x}(t,\theta)$, following (1.9.39).

3. Calculate $\hat{\nu}^x_j(u,\boldsymbol{\theta})$, $\hat{\eta}_{s,d}(W,x,t,\theta)$, and $\hat{\psi}^x_j(W,\theta)$, based on (1.5.1) and (1.5.6) - (1.5.9), respectively.

4. Sample $\{\xi^b_i\}^n_{i=1}$ from a distribution with zero mean and unit variance, independently from data. Calculate $\hat{\boldsymbol{\psi}}^{x*}_j$, and $\mathbb{G}^x_{\xi^b,j}(u,\theta)$.

   Repeat Step 4 for $b=1,...,B$, where B is some large integer.

5. For $\ell = lb, ub$, compute the $(1-\alpha)$-th quantile $\hat{c}^{x,B}_{n,\ell,j}(\alpha, \mathscr{U}_m, \mathbf{\Theta}_l)$ of $\left\{\max_{1\leq i\leq m, 1\leq s\leq l} \left\| \mathbb{G}^x_{\ell,\xi^b,j}(u_i, \boldsymbol{\theta}) \right\| \right\}_{b=1}^{B}$, and construct the uniform confidence band

$$C^{x,B}_{n,\ell,j}(1-\alpha, \mathscr{U}_m, \mathbf{\Theta}_l) \equiv \left\{ \hat{v}^x_{\ell,j}(u, \boldsymbol{\theta}) \pm n^{-1/2} h^{-1/2} \hat{c}^{x,B}_{n,\ell,j}(\alpha, \mathscr{U}_m, \mathbf{\Theta}_l) : u \in \mathscr{U}_m, \boldsymbol{\theta} \in \mathbf{\Theta}_l \right\}.$$

Proof of Theorem 1.6. The proof is a direct consequence of Theorem 1.4, Corollary 1.3, and the continuous mapping theorem for the multiplier bootstrap, cf. Theorem 2.6 in Kosorok (2008). ■

1.9.2 Single-Index Estimator

In this section, we establish large sample properties of the index coefficients estimator $\hat{\boldsymbol{\gamma}}$. The results, as presented in the following lemma, are largely based on Proposition 1 in Li and Patilea (2018). We show that $\hat{\boldsymbol{\gamma}}$ is consistent for $\boldsymbol{\gamma}$, converges to $\boldsymbol{\gamma}$ at the parametric rate. Moreover, $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ admits an asymptotic linear representation, and converges in distribution to a normal distribution.

Lemma 1.3 Under Assumptions 1.1, 1.3, 1.4.1, 1.5, 1.6.1–1.6.3, 1.7, 1.8.1 and 1.9.1, it holds that

$$\hat{\gamma}_d - \gamma_d = \frac{1}{n} \sum_{i=1}^{n} V_d^{-1} \psi_d^b(W_i) + o_p\left(n^{-1/2}\right), \tag{1.9.32}$$

for $d \in \{0,1\}$, where ψ_d^a and V_d are defined in Section 1.4.3. Furthermore,

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{d} N(0, \Sigma_{\boldsymbol{\gamma}}),$$

where

$$\Sigma_{\boldsymbol{\gamma}} \equiv \begin{pmatrix} \Sigma_{\gamma_1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\gamma_0} \end{pmatrix},$$

and $\Sigma_{\gamma_d} \equiv V_d^{-1} \mathbb{E}[\psi_d^b(W) \psi_d^b(W)'] V_d^{-1}$.

Proof of Lemma 1.3. The linear expansion follows from the same lines of argument as in that of Proposition 1 in Li and Patilea (2018). We show how their Assumption 8.1 can be fulfilled by parallel conditions in our context. Condition (1) is satisfied under Assumptions 1.5.1 and 1.3.1. Condition (2) holds under Assumptions 1.4.1 and 1.5.2. Conditions (3) and (4) follow from Theorem 1.1, which holds under Assumptions 1.1–1.3. Condition (5) is satisfied under 1.3.1.

Condition (6) is due to Assumptions 1.6.1–1.6.3. Lastly, Assumptions 1.7.2, 1.7.1, 1.8.1 imply Conditions (7)–(9), respectively. We remark that Assumption 1.9.1 is slightly weaker than Condition (10). However, their proof carries through, under this weaker condition, if the maximal inequality from Lemma 1.5, instead of the main corollary of Sherman (1994), is employed in the proof.

To conclude, we note that $\psi_1^b \cdot \psi_0^b = 0$, and thus the covariance between $\hat{\gamma}_1 - \gamma_1$ and $\hat{\gamma}_0 - \gamma_0$ is 0. ■

1.9.2.1 Single-Index Kernel Estimator

In this section, we present a lemma documenting some well-known facts on single-index kernel estimators. These results will be used repeated throughout the appendix. First, we introduce some quantities,

$$\begin{aligned}
&K^{(1)}(u) \equiv dK(u)/du, \quad K^{(2)}(u) \equiv d^2K(u)/du^2, \quad K_h^{(j)}(u,v) \equiv K^{(j)}((u-v)/h)/h, \text{ for } j = 1,2, \\
&G_d^{(1)}(y,x\gamma) \equiv \left\{\rho_{1,1}^{\gamma}/\rho_{0,0}^{\gamma} - \rho_{1,0}^{\gamma}\rho_{0,1}^{\gamma}/\rho_{0,0}^{\gamma\ 2}\right\}(y,x\gamma), \qquad (1.9.33)
\end{aligned}$$

and analogous definition for the sub-distributions should be apparent. We remark that $G_d^{(1)}(y,x\gamma)$ is, in general, not equal to $\partial_\gamma G_d(y,x\gamma)$. When $\gamma = \gamma_d$, the expression simplifies, and we find, by direct calculations, that $G_d^{(1)}(y,x\gamma_d) = \partial_{x\gamma} G_d(y,x\gamma_d) \cdot \mathbb{E}\left[(x-X)|X\gamma_d = x\gamma_d\right]$.

In the following lemma, we provide convergence rates for single-index kernel density and conditional distribution estimators. Let $\mathscr{H} \equiv [h_l n^{-\zeta}, h_u]$, for some positive number h_l, h_u, and $0 < \zeta < 1/3$.

Lemma 1.4 Suppose Assumptions 1.3.1, 1.5, 1.4.1, 1.6.3, 1.8.2, and 1.9.2 hold, then

$$\sup_{(y,x,h,\gamma)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\mathscr{H}\times\Gamma} \left|\hat{f}_d(x\gamma) - \rho_{0,0}(y,x\gamma)\right| = O_p\left((\log n)^{1/2} n^{-1/2} h^{-1/2}\right) + O(h^s), \qquad (1.9.34)$$

$$\sup_{(y,x,h,\gamma)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\mathscr{H}\times\Gamma} \left|\hat{\kappa}_{d,y}(x\gamma) - \rho_{1,0}(y,x\gamma)\right| = O_p\left((\log n)^{1/2} n^{-1/2} h^{-1/2}\right) + O(h^s), \qquad (1.9.35)$$

$$\sup_{(y,x,h,\gamma)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\mathscr{H}\times\Gamma} \left|\hat{G}_d(y,x\gamma) - G_d(y,x\gamma)\right| = O_p\left((\log n)^{1/2} n^{-1/2} h^{-1/2}\right) + O(h^s) \qquad (1.9.36)$$

$$\sup_{(y,x,h,\gamma)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\mathscr{H}\times\Gamma} \left\|\partial_\gamma \hat{G}_d(y,x\gamma) - G_d^{(1)}(y,x\gamma)\right\| = O_p\left((\log n)^{1/2} n^{-1/2} h^{-3/2}\right) + O(h^s), \qquad (1.9.37)$$

for $\ell = 1,2$. Analogous results hold for $\hat{G}_{d,r}$, and $\partial_\gamma \hat{G}_{d,r}$, with $r = 0,1$.

Proof of Lemma 1.4. The proof follows from standard kernel techniques, cf. Einmahl and Mason (2005). It is also implicit in that of Theorem 1 in Chiang and Huang (2012), and hence, is omitted. ■

1.9.2.2 Test of Single-Index Assumption

In this section, we introduce a formal specification test of Assumption 1.2. Let

$$
\begin{aligned}
\hat{\mathscr{J}}_d(\gamma;\rho,g) &= \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\rho(X_i'Q(\gamma),X_j'Q(\gamma))J_g(X_i'\gamma,X_j'\gamma)\sum_{\ell=1}^{n}\hat{U}_{d,\gamma,i}(Y_\ell,R_\ell)\hat{U}_{d,\gamma,j}(Y_\ell,R_\ell)\right\},\\
\hat{V}_{\mathscr{J}_d}(\gamma;\rho,g) &= \frac{1}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\rho(X_i'Q(\gamma),X_j'Q(\gamma))^2J_g(X_i'\gamma,X_j'\gamma)^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\cdot\left\{\frac{1}{n}\sum_{\ell=1}^{n}\hat{U}_{d,\gamma,i}(Y_\ell,R_\ell)\hat{U}_{d,\gamma,j}(Y_\ell,R_\ell)\right\}^2,\\
\hat{T}_{\mathscr{J}_d}(\gamma;\rho,g) &= \frac{\hat{\mathscr{J}}_d(\gamma;\rho,g)}{\sqrt{\hat{V}_{\mathscr{J}_d}(\gamma;\rho,g)}},\\
\hat{\mathscr{J}}_d(\gamma,\xi;\rho,g) &= \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\xi_i\xi_j\left\{\rho(X_i'Q(\gamma),X_j'Q(\gamma))J_g(X_i'\gamma,X_j'\gamma)\sum_{\ell=1}^{n}\hat{U}_{d,\gamma,i}(Y_\ell,R_\ell)\hat{U}_{d,\gamma,j}(Y_\ell,R_\ell)\right\},\\
\hat{V}_{\mathscr{J}_d}(\gamma,\xi;\rho,g) &= \frac{1}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\xi_i^2\xi_j^2\rho(X_i'Q(\gamma),X_j'Q(\gamma))^2J_g(X_i'\gamma,X_j'\gamma)^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\cdot\left\{\frac{1}{n}\sum_{\ell=1}^{n}\hat{U}_{d,\gamma,i}(Y_\ell,R_\ell)\hat{U}_{d,\gamma,j}(Y_\ell,R_\ell)\right\}^2,\\
\hat{T}_{\mathscr{J}_d}(\gamma,\xi;\rho,g) &= \frac{\hat{\mathscr{J}}_d(\gamma,\xi;\rho,g)}{\sqrt{\hat{V}_{\mathscr{J}_d}(\gamma,\xi;\rho,g)}}.
\end{aligned}
$$

Algorithm 1.9.2 1. Solve for the minimizer of $\hat{\mathscr{J}}_d(\gamma;\rho)$ in $\gamma\in\Gamma$, $\hat{\gamma}_d$, for each $d\in\{0,1\}$.

2. Search for a $(k-1)\times(k-2)$ matrix $Q_{-1}(\hat{\gamma}_d)$ such that $Q(\hat{\gamma}_d)=[\hat{\gamma}_d\ \ Q_{-1}(\hat{\gamma}_d)]$ is an orthogonal matrix.

3. Calculate $\hat{\mathscr{J}}_d(\hat{\gamma}_d;\rho,g)$, $\hat{V}_{\mathscr{J}_d}(\hat{\gamma}_d;\rho,g)$, and $\hat{T}_{\mathscr{J}_d}(\hat{\gamma}_d;\rho,g)$.

4. Sample $\{\xi_i^b\}_{i=1}^n$ from a distribution with zero mean and unit variance, independently from data. Calculate $\hat{\mathscr{J}}_d(\hat{\gamma}_d,\xi^b;\rho,g)$, $\hat{V}_{\mathscr{J}_d}(\hat{\gamma}_d,\xi^b;\rho,g)$, and $\hat{T}_{\mathscr{J}_d}(\hat{\gamma}_d,\xi^b;\rho,g)$.

   Repeat Step 4 for $b=1,...,B$, where B is some large integer.

5. Compute the $\alpha/2$-th and $(1-\alpha/2)$-th quantiles of $\{\hat{T}_{\mathscr{J}_d}(\hat{\gamma}_d,\xi^b;\rho,g)\}_{b=1}^B$, reject the null hypothesis that Assumption 1.2 holds if $\hat{T}_{\mathscr{J}_d}(\hat{\gamma}_d;\rho,g)$ lies outside the interval defined by these two critical values.

1.9.3 Auxiliary Results

1.9.3.1 Definitions and Additional Results

In this section, we provide several results related to the results in the main text. First, we introduce the notion of *covering numbers* and the *VC type* (or *Euclidean*) class. Let $\|\cdot\|_{L^r(P)}$ denote $\{\mathbb{E}[|f(W)|^r]\}^{1/r}$ Given a class of functions

$\mathscr{F}$ defined on a space $\mathscr{X}$, a probability measure Q, the covering number $\mathscr{N}(\varepsilon, \mathscr{F}, L_r(Q))$, is the minimum number of $L_r(Q)$ balls of radius ε needed to cover $\mathscr{F}$. The centers of these balls is not required to be in $\mathscr{F}$. A function $F : \mathscr{X} \mapsto \mathbb{R}$ is called an envelop for $\mathscr{F}$ if $|f(x)| \leq F(x)$ for all $x \in \mathscr{X}$ and all $f \in \mathscr{F}$. We say $\mathscr{F}$ is of the VC type with respect to an envelop function F if there exists positive constants A and $V \geq 1$ satisfying $\sup_Q \mathscr{N}(\varepsilon \|F\|_{L_2(Q)}, \mathscr{F}, L_2(Q)) \leq (A \|F\|_{L_2(Q)} / \varepsilon)^V$, for all $\varepsilon \in (0, 2\|F\|_{L_2(Q)}]$ and the supremum is taken over all probability measures Q with $0 < \|F\|_{L_2(Q)} < \infty$. We say that the VC class $\mathscr{F}$ admits the characteristics A and V.

Given a function g of m variables, let $U_n^{(m)}(g) = \frac{(n-k)!}{n!} \sum_{i \in I_n^m} g(X_{i_1}, ..., X_{i_m})$, with $I_n^m = \{(i_1, ..., i_k) : 1 \leq i_j \leq n, i_j \neq i_l, if j \neq l\}$. Let the *Hoeffding projections* of g with respect to a measure P be defined as $\pi_k g = (\delta_{x_1} - P) \times \cdots \times (\delta_{x_k} - P) \times P^{m-k} g$ and $\pi_0 = \mathbb{E}[g(X_1, ..., X_m)]$. If g is symmetric in its entries, we define the *Hoeffding decomposition* as $U_n^{(m)}(g) - \mathbb{E}[g] = \sum_{j=1}^m \frac{m!}{j!(m-j)!} U_n^{(j)}(\pi_j g)$. The following Lemma, due to Giné and Mason (2007), establishes a maximal inequality for moment of the U processes which plays a crucial role in our derivation of the uniform linear representations.

Lemma 1.5 (Giné and Mason, 2007, Theorem 8) Let $\mathscr{F}$ be a measurable collection of symmetric functions $S^m \mapsto \mathbb{R}$ with an envelop function F and let P be any probability measure on the space $(S, \mathscr{S})$. Assume F is bounded by $M > 0$ and $\mathscr{F}$ is a VC class with respect to F with characteristics A and V. Then for every $m \in \mathbb{N}$, $A \geq e^m$, $V \geq 1$, there exist constants C_1 and C_2 such that

$$n^{k/2} \mathbb{E}\left[\left\|U_n^{(k)}(\pi_k f)\right\|_{\mathscr{F}}\right] \leq C_1 \sigma \left(\log \frac{A \|F\|_{L_2(P^m)}}{\sigma}\right)^{k/2},$$

for $k = 0, 1, ..., m$, assuming $n\sigma^2 \leq C_2 \log\left(2 \|F\|_{L_2(P^m)} / \sigma\right)$, where σ^2 satisfies $\left\|P^m f^2\right\| \leq \sigma^2 \leq P^m F^2$.

Let inner and outer expectations be denoted by $\mathbb{E}_*$ and $\mathbb{E}^*$ as in Section 1.2 of Van Der Vaart and Wellner (1996). We say a sequence of stochastic process $X_n : \mathbb{E} \mapsto \mathbb{D}$, where $\mathbb{E}$ and $\mathbb{D}$ are metric spaces, *converges weakly* to X, denoted by $X_n \Rightarrow X$ if $\mathbb{E}^*[h(X_n)] \to \mathbb{E}[h(X)]$, for all $h \in \mathscr{C}_b(\mathbb{D})$ where $\mathscr{C}_b(\mathbb{D})$ denotes the space of the real-valued bounded continuous functions defined on $\mathbb{D}$.

We follow the definition of *conditional weak convergence in probability* as appeared in Section 2.2.3 in Kosorok (2008). The notation $X_n \underset{\xi}{\overset{p}{\to}} X$ means that $\sup_{h \in BL_1} \left|\mathbb{E}_{\xi|w}[h(X_n)] - \mathbb{E}[h(X)]\right| \xrightarrow{p} 0$ and $\mathbb{E}^*_{\xi|w}[h(X_n)] - \mathbb{E}_{\xi|w_*}[h(X_n)] = 0$, where BL_1 is the space of functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1. Namely, $\|f\|_\infty \leq 1$ and $|f(x) - f(y)| \leq d(x, y)$, for $x, y \in \mathbb{D}$. The operator $\mathbb{E}_{\xi|w}$ denotes the conditional expectation over the weights ξ given the remaining data.

The following lemma, originated from Theorem 10.6 of Pollard (1990) and restated in Theorem 11.16 of Kosorok (2008) is key to establishing weak convergence of conditional processes.

Lemma 1.6 (Kosorok, 2008, Theorem 11.16) Suppose a triangular array stochastic processes $\{f_{ni}(t) : i = 1,...,n, t \in T\}$ consisting of row-wise independent processes is almost measurable Suslin (AMS). Define $\chi_n(t) = \sum_{i=1}^n f_{ni}(t)$ and $\rho_n(s,t) = \left(\sum_{i=1}^n \mathbb{E}\left[(f_{ni}(s) - f_{ni}(t))^2\right]\right)^{1/2}$, for $s,t \in T$. (i) the $\{f_{ni}\}$ are manageable, with envelops $\{F_{ni}\}$ which are also independent within rows; (ii) $H(s,t) = lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}[\chi_n(s)\chi_n(t)]$ exists for every $s,t \in T$; (iii) $\limsup_{n\to\infty} \mathbb{E}^*[F_{ni}^2] < \infty$; (iv) $\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}^*\left[F_{ni}^2 \mathbb{1}\{F_{ni}\} > \varepsilon\right] = 0$; (v) $\rho(s,t) = \lim_{n\to\infty} \rho_n(s,t)$ exists for every $s,t \in T$. For all deterministic sequences $\{s_n\}$ and $\{t_n\}$, if $\rho(s_n,t_n) \to 0$ then $\rho_n(s_n,t_n) \to 0$. Then T is totally bounded under the ρ pseudo-metric and χ_n converges weakly on $\ell_\infty(T)$ to a tight, centered Gaussian process χ concentrated on $\{g \in \ell_\infty(T) :$ g is uniformly ρ-continuous$\}$, with covariance function $H(s,t)$.

Precise definitions of AMS and manageable triangular arrays can be found in Section 11.4.1 in Kosorok (2008). Direct check of these two conditions is usually not easy. To address this issue, Kosorok (2008) presents sufficient conditions: by Lemma 11.15 in Kosorok (2008), the triangular array is AMS whenever it is separable[12], and for manageability to hold, Lemma 11.21 in Kosorok (2008) implies that the VC type condition on the triangular array suffices.

Lastly, we recall the definition of *Hadamard differentiability*, see pp. 272-273 in Van Der Vaart and Wellner (1996). We say a mapping $\nu : \mathbb{D}_\nu \subset \mathbb{D} \to \mathbb{E}$ is called *Hadamard differentiable* at $F \in \mathbb{D}_\nu$, tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$, if there is a continuous linear map $\nu_F^{'} : \mathbb{D} \to \mathbb{E}$ such that

$$\frac{\nu(F + t_n h_n) - \nu(F)}{t_n} \to \nu_F^{'}(h),$$

for all converging sequences $\{t_n\} \subset \mathbb{R}$ with $t_n \to 0$ and $\{h_n\} \subset \mathbb{D}$ with $h_n \to h \in \mathbb{D}_0$, such that $F + t_n h_n \in \mathbb{D}_\nu$ as $n \to \infty$, for all n.

1.9.3.2 Auxiliary Lemmas

Lemma 1.7 Suppose that the assumptions of Theorem 1.3 hold. The function classes, $\mathscr{G}_1$ - $\mathscr{G}_6$, $\mathscr{G}_b$, $\mathscr{G}_\eta$, and $\mathscr{G}_\varphi$ as defined in (1.9.19), (1.9.20),(1.9.22), (1.9.23), (1.9.27), (1.9.28), (1.9.24), (1.9.25), and (1.9.29) are of VC type with bounded envelop.

Proof of Lemma 1.7. We first identify the sub-classes that constitute the above functional classes and show that the uniform entropy condition is satisfied for each of these sub-classes. Then, we illustrate on how we use results on the

[12] A triangular array of stochastic process $\{f_{ni}(t) : i = 1,...,n, t \in T\}$ is separable if, for all $n \geq 1$, there exists a countable set $T_n \in T$ such that

$$\mathbb{E}^*\left[\mathbb{1}\left\{\sup_{t\in T}\sup_{s\in T_n}\sum_{i=1}^n (f_{ni}(s) - f_{ni}(t))^2 > 0\right\}\right] = 0.$$

sub-classes to show that the functional classes in the theorem is of VC type. Define

$$\begin{aligned}
\mathscr{M}_1 &\equiv \{y \mapsto \mathbb{1}\{y \leq t\} : t \in \tilde{\mathscr{T}}\},\\
\mathscr{M}_2 &\equiv \{x_1 \mapsto K((x_1\gamma_d - x\gamma_d)/h)\mathbb{1}\{|x_1\gamma_d - x\gamma_d| \leq h\} : (x,h) \in \mathscr{X} \times \mathscr{H}\},\\
\mathscr{M}_3 &\equiv \{x_1 \mapsto K^{(1)}((x_1\gamma_d - x\gamma_d)/h)\mathbb{1}\{|x_1\gamma_d - x\gamma_d| \leq h\} : (x,h) \in \mathscr{X} \times \mathscr{H}\},\\
\mathscr{M}_{4,1} &\equiv \{y \mapsto \partial^{\ell}_{x\gamma} G_d(y, x\gamma_d) : \ell \in \{0,1,2,...,s\}, (d,x) \in \{0,1\} \times \mathscr{X}\},\\
\mathscr{M}_{4,2} &\equiv \{y \mapsto \partial^{\ell}_{x\gamma} G_{d,1}(y, x\gamma_d) : \ell \in \{0,1,2,...,s\}, (d,x) \in \{0,1\} \times \mathscr{X}\}.\\
\mathscr{M}_{4,3} &\equiv \{\partial G_d(y, x\gamma_d)/\partial y|_{y=t} : (d,t,x) \in \{0,1\} \times \tilde{\mathscr{T}} \times \mathscr{X}\}.\\
\mathscr{M}_{4,4} &\equiv \{\partial G_{d,1}(y, x\gamma_d)/\partial y|_{y=t} : (d,t,x) \in \{0,1\} \times \tilde{\mathscr{T}} \times \mathscr{X}\}.\\
\mathscr{M}_5 &\equiv \{\partial^{\ell}_{x\gamma} f_d(x\gamma_d) : \ell \in \{0,1,2,...,s\}, (d,x) \in \{0,1\} \times \mathscr{X}\}.
\end{aligned}$$

By Lemma 19.15 in Van der Vaart (1998), $\mathscr{M}_1$ is of VC type with the constant envelop. Under Assumption 1.8.2, both $K(\cdot)$ and $K^{(1)}(\cdot)$ are of bounded variation, Lemma 22(i) of Nolan and Pollard (1987) implies that $\mathscr{M}_{2,1}$ and $\mathscr{M}_{2,2}$ belong to the VC class with a constant envelop. Next, since $\partial^{\ell}_{v} F_{Y_d,R_d|D,X\gamma_d}(y,r|d,v)$, $\ell = 0,...,s$, is Lipschitz continuous with respect to $x\gamma_d$ under Assumption 1.6.2(i), Lemma 2.13 of Pakes and Pollard (1989) implies $\mathscr{M}_{4,1}$ and $\mathscr{M}_{4,2}$ are of VC type with bounded envelop functions. The proof for $\mathscr{M}_{4,3}, \mathscr{M}_{4,4}$, and $\mathscr{M}_5$ follows the same arguments based on the Lipschitz continuity of $\partial_y F_{Y_d,R_d|D,X\gamma_d}(y,r|d,v)$ with respect to y and v, and of $\partial_v f_{d,\gamma_d}(v)$ with respect to v, as implied by Assumption 1.6.2(iv), and 1.6.1(i), respectively.

Now we are ready to show why the functional classes in the lemma are of VC type. We illustrate on $\mathscr{G}_1$ and $\mathscr{G}_\eta$. All others follow by same lines of reasoning.

We focus on $\mathscr{G}_1$ first. Note that the class that g_{11} belongs to is a product of a finite set $\{(r_1,d_1) \mapsto r_1\mathbb{1}\{d_1 = d\}, d \in \{0,1\}\}$, $\mathscr{M}_1$, and $\mathscr{M}_2$, and thus, it is of VC type by Corollary A.1 in Chernozhukov et al. (2014). Since all three sub-classes have finite envelops, their product also does. Regarding g_{12}, we first show that $\mathscr{M}_\phi \equiv \{y \mapsto \phi''_\theta(s_d(y, x\gamma_d)) : (x,\theta) \in \mathscr{X} \times \Theta\}$ is also a VC class with bounded envelop. For any $x_1, x_2 \in \mathscr{X}$ and $\theta_1, \theta_2 \in \Theta$, we have

$$\begin{aligned}
&\left|\phi''_{\theta_1}(s_d(y, x_1\gamma_d)) - \phi''_{\theta_2}(s_d(y, x_2\gamma_d))\right|\\
\leq &\left|\phi''_{\theta_1}(s_d(y, x_1\gamma_d)) - \phi''_{\theta_2}(s_d(y, x_1\gamma_d))\right| + \left|\phi''_{\theta_2}(s_d(y, x_1\gamma_d)) - \phi''_{\theta_2}(s_d(y, x_2\gamma_d))\right|\\
\leq &M_1 |\theta_1 - \theta_2| + \sup_{(\theta,u) \in \Theta \times [\upsilon_o, 1]} \left|\phi'''_\theta(u)\right| \sup_{(y,x) \in \tilde{\mathscr{T}} \times \mathscr{X}} \left|\partial_{x\gamma} G_d(y, x\gamma_d)\right| \|x_1 - x_2\| \|\gamma_d\|\\
\leq &M_1 |\theta_1 - \theta_2| + M_2 \|x_1 - x_2\| \leq \sqrt{2}\max\{M_1, M_2\} \left\|(\theta_1, x'_1)' - (\theta_2, x'_2)'\right\|,
\end{aligned}$$

where M_1 and M_2 are positive constants. The second inequality is due to the Lipschitz continuity condition on $\phi_\theta^{''}$, and the third follows because $\phi_\theta^{'''}$ and $\partial_{x\gamma}G_d$ are uniformly bounded under Assumption 1.6.4, and 1.6.2. The last one is by Hölder's inequality. Another application of Lemma 2.13 of Pakes and Pollard (1989) yields the desired result.

Let $\mathscr{M}_x = \{\tilde{x} \mapsto \tilde{x}_\ell - x_\ell : \ell = 2, ..., k, x \in \mathscr{X}\}$. Since $\mathscr{X}$ is compact, $\mathscr{M}_x$ is a VC class because $\mathscr{N}(\varepsilon \sup_{x\in\mathscr{X}} \left\|x_{[-1]}\right\|, \mathscr{M}_x, L_2(Q)) \leq C(diam(\mathscr{X})/\varepsilon)$, for a positive constant C independent of ε. Applying Corollary A.1 in Chernozhukov et al. (2014) again on the product of $\mathscr{M}_1$, $\mathscr{M}_3$, $\mathscr{M}_\phi$, $\mathscr{M}_x$, and the finite set $\{(y_1, y_2, d_2) \mapsto \mathbb{1}\{d_2 = d, y_2 \leq y_1\}, d \in \{0,1\}\}$ yields that the first half of g_{12} belongs to a VC class. Next, by Lemma 5 of Sherman (1994), we deduce that $\{y_1 \mapsto \int \mathbb{1}\{d_2 = d, y_2 \leq y_1 \wedge t\} hK^{(1)}(x\gamma_d, x_2'\gamma_d)(x_{2,l} - x_l)dF(w_2) : \omega \in \Omega\}$ and $\{w_2 \mapsto \int g_{11}(w_1, \omega) g_{12}(w_2, y_1, \omega) dF(w_1) : \omega \in \Omega\}$ are both of the VC type. Since $f_d(x\gamma_d)$ is uniformly bounded away from 0, $\{1/f_d(x\gamma_d)^2 : (d,x) \in \{0,1\} \times \mathscr{X}\}$ admits a finite envelop. Applying Corollary A.1 in Chernozhukov et al. (2014) yet again concludes the proof.

Turning to $\mathscr{G}_\eta$, we first show that for a fixed x, the set $\mathscr{M}_\theta \equiv \{1/\phi_\theta^{'}(s_{T_d}(t, x\gamma_d, \theta)) : (t, \theta) \times \tilde{\mathscr{T}} \times \Theta\}$ belongs to the VC class. Recall that $s_{T_d}(t, x\gamma_d, \theta) = \phi_\theta^{-1}\left(\int_0^t \phi_\theta^{'}(s_d(y, x\gamma_d)) s_{d,1}(dy, x\gamma_d)\right)$. Hence,

$$\begin{aligned}
&1/\phi_{\theta_1}^{'}(s_{T_d}(t_1, x\gamma_d, \theta_1)) - 1/\phi_{\theta_2}^{'}(s_{T_d}(t_2, x\gamma_d, \theta_2)) \\
\leq &\left\{1/\phi_{\theta_1}^{'}(s_{T_d}(t_1, x\gamma_d, \theta_1)) - 1/\phi_{\theta_2}^{'}(s_{T_d}(t_1, x\gamma_d, \theta_2))\right\} \\
&+ \left\{1/\phi_{\theta_2}^{'}(s_{T_d}(t_1, x\gamma_d, \theta_2)) - 1/\phi_{\theta_2}^{'}(s_{T_d}(t_2, x\gamma_d, \theta_2))\right\} \\
\equiv &\Delta_1 + \Delta_2.
\end{aligned}$$

Decomposing the first term further into,

$$\begin{aligned}
|\Delta_1| \leq &\left|1/\dot{\phi}_{\theta_1}^{-1}\left(\int_0^{t_1} \phi_{\theta_1}^{'}(s_d(y, x\gamma_d)) s_{d,1}(dy, x\gamma_d)\right) - 1/\dot{\phi}_{\theta_1}^{-1}\left(\int_0^{t_1} \phi_{\theta_2}^{'}(s_d(y, x\gamma_d)) s_{d,1}(dy, x\gamma_d)\right)\right| \\
&+ \left|1/\dot{\phi}_{\theta_1}^{-1}\left(\int_0^{t_1} \phi_{\theta_2}^{'}(s_d(y, x\gamma_d)) s_{d,1}(dy, x\gamma_d)\right) - 1/\dot{\phi}_{\theta_2}^{-1}\left(\int_0^{t_1} \phi_{\theta_2}^{'}(s_d(y, x\gamma_d)) s_{d,1}(dy, x\gamma_d)\right)\right| \\
\equiv &\Delta_{11} + \Delta_{12}.
\end{aligned}$$

For the first term, we have

$$|\Delta_{11}| \leq (1 - \upsilon_o) \sup_{(z,\theta)\in[0,y_o^*]\times\Theta} \left|\frac{\phi_\theta^{''}(\phi_\theta^{-1}(z))}{\left(\dot{\phi}_\theta^{-1}(z)\right)^3}\right| \sup_{(u,\theta)\in[\upsilon_o,1]\times\Theta} \left|\phi_\theta^{'}(u)\right| |\theta_1 - \theta_2| = M_3 |\theta_1 - \theta_2|.$$

Under Assumption 1.10.(ii), $\Delta_{12} \leq M_4 \left|\theta_1 - \theta_2\right|$.

$$\begin{aligned}
|\Delta_2| \leq & \sup_{(z,\theta)\in[0,y_o^*]\times\Theta} \left| \frac{\ddot{\phi}_\theta^{-1}(z)}{\left(\dot{\phi}_\theta^{-1}(z)\right)^3} \right| \\
& \cdot \left| \int_0^{t_1} \phi_{\theta_2}^{'}(s_d(y,x\gamma_d))s_{d,1}(dy,x\gamma_d) - \int_0^{t_2} \phi_{\theta_2}^{'}(s_d(y,x\gamma_d))s_{d,1}(dy,x\gamma_d) \right| \\
\leq & \sup_{(z,\theta)\in[0,y_o^*]\times\Theta} \left| \frac{\ddot{\phi}_\theta^{-1}(z)}{\left(\dot{\phi}_\theta^{-1}(z)\right)^3} \right| \cdot \sup_{(u,\theta)\in[\upsilon_o,1]\times\Theta} \left|\phi_\theta^{'}(u)\right| \sup_{(y,x)\in\tilde{\mathscr{T}}\times\mathscr{X}} \left|\partial_{x\gamma} G_{d,1}(y,x\gamma_d)\right| |t_1 - t_2| \\
= & M_5 \left|t_1 - t_2\right|.
\end{aligned}$$

where inequalities hold by the mean value theorem and under Assumptions 1.6.2, and 1.6.4. Combining the bounds and applying Hölder's inequality, we conclude by Lemma 2.13 of Pakes and Pollard (1989) that $\mathscr{M}_\theta$ is a VC class.

Next, following similar analysis as in the previous part, we deduce from Corollary A.1 of Chernozhukov et al. (2014) that $\{w_1 \mapsto \mathbb{1}\{d_1 = d\}(\mathbb{1}\{y_1 \leq y\} - G_d(y,x_1\gamma_d))\,\partial_y G_{d,1}(y,x\gamma_d) : (d,y,\theta) \in \{0,1\} \times \tilde{\mathscr{T}} \times \Theta\}$ for a given $x \in \mathscr{X}$ is a VC class with a finite envelop. Applying Lemma 5 of Sherman (1994), we get $\{w_1 \mapsto \int \mathbb{1}\{y_1 \leq t\}\mathbb{1}\{d_1 = d\} \cdot (\mathbb{1}\{y_1 \leq y\} - G_d(y,x_1\gamma_d))\,\partial_y G_{d,1}(y,x\gamma_d)dy : (d,t,\theta) \in \{0,1\} \times \tilde{\mathscr{T}} \times \Theta\}$ also belongs to the VC class with an envelop $F_{\eta,1} = G_{d,1}(y_o \wedge y_c, x\gamma_d)$. This is due to

$$\begin{aligned}
&\int \mathbb{1}\{y_1 \leq t\}\mathbb{1}\{d_1 = d\}(\mathbb{1}\{y_1 \leq y\} - G_d(y,x_1\gamma_d))\,\partial_y G_{d,1}(y,x\gamma_d)dy \\
&\quad \leq 2\int_0^t \partial_y G_{d,1}(y,x\gamma_d)dy \leq G_{d,1}(y_o \wedge y_c, x\gamma_d).
\end{aligned}$$

Analogous results can be established for the other two parts of Ψ_d.

Combining these results with the fact that $\{x_1 \mapsto K((x_1\gamma_d - x\gamma_d)/h) : (t,h) \in \tilde{\mathscr{T}} \times \mathscr{H}\}$ is VC with an envelop $C\mathbb{1}\{|x_1\gamma_d - x\gamma_d| \leq h\}$, we deduce that $\mathscr{G}_\eta$ is of the VC type, with the envelop given by $\sum_{d=0,1} C_d \mathbb{1}\{|x_1\gamma_d - x\gamma_d| \leq h\}$ where C_0 and C_1 are positive constants. Setting $H_{\eta,d}(x_1\gamma_d) = C_d\mathbb{1}\{|x_1\gamma_d - x\gamma_d| \leq h\}$ concludes the proof. ■

Lemma 1.8 Under the assumptions of Theorem 1.3, for any $\delta_n = O_p\left(n^{-1/2}\right)$,

$$\begin{aligned}
\sup_{\|\tilde{\gamma}_d - \gamma_d\| \leq \delta_n} \sup_{(t,x)\in\tilde{\mathscr{T}}\times\mathscr{X}} & \left\| \int_0^t \ddot{\phi}_{d,\gamma_d}^\theta(y,x)\left\{\partial_\gamma \hat{G}_d(y,x\tilde{\gamma}_d) - \partial_\gamma \hat{G}_d(y,x\gamma_d)\right\} s_{d,1}(dy,x\gamma_d) \right\| \\
& = O_p\left((\log n)^{1/2} n^{-1/2} h^{-5/2} \delta_n\right) + O(\delta_n).
\end{aligned}$$

Proof of Lemma 1.8. Split the term inside the norm operator into

$$\begin{aligned}\Delta_1(t,x,\theta) &\equiv \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{\frac{\partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d)}{\hat{f}_d(x\tilde{\gamma}_d)} - \frac{\partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)}{\hat{f}_d(x\gamma_d)}\right\} s_{d,1}(dy,x\gamma_d),\\ \Delta_2(t,x,\theta) &\equiv -\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{\frac{\hat{\kappa}_{d,y}(x\tilde{\gamma}_d)\partial_\gamma \hat{f}_d(x\tilde{\gamma}_d)}{\hat{f}^2_d(x\tilde{\gamma}_d)} - \frac{\hat{\kappa}_{d,y}(x\gamma_d)\partial_\gamma \hat{f}_d(x\gamma_d)}{\hat{f}^2_d(x\gamma_d)}\right\} s_{d,1}(dy,x\gamma_d).\end{aligned}$$

Decomposing Δ_1 and ignoring smaller order terms gives

$$\begin{aligned}\Delta_1(t,x,\theta) =& f_d(x\gamma_d)^{-1}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\left\{\partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d) - \partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)\right\} s_{d,1}(dy,x\gamma_d)\\ &+ \frac{\hat{f}_d(x\tilde{\gamma}_d) - \hat{f}_d(x\gamma_d)}{f_d(x\tilde{\gamma}_d)f_d(x\gamma_d)}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)s_{d,1}(dy,x\gamma_d) + (s.o.).\end{aligned}$$

We investigate the uniform rate of the first term only. The second term exhibits the same rate and is simpler. Define

$$\begin{aligned}\Delta_{11}(t,x,\theta,\tilde{\gamma}_d) &\equiv \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\mathbb{E}[\partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d) - \partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)]s_{d,1}(dy,x\gamma_d),\\ \Delta_{12}(t,x,\theta,\tilde{\gamma}_d) &\equiv \int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\{\partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d) - \partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)\\ &\qquad - \mathbb{E}[\partial_\gamma \hat{\kappa}_{d,y}(x\tilde{\gamma}_d) - \partial_\gamma \hat{\kappa}_{d,y}(x\gamma_d)]\}s_{d,1}(dy,x\gamma_d).\end{aligned}$$

By Fubini's theorem and standard change of variables,

$$\begin{aligned}\Delta_{11}(t,x,\theta,\tilde{\gamma}_d) = & h^{-2}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\cdot\\ &\mathbb{E}\left[\rho^{\tilde{\gamma}_d}_{1,1}(y,X\tilde{\gamma}_d)K^{(1)}((X\tilde{\gamma}_d - x\tilde{\gamma}_d)/h) - \rho^{\gamma_d}_{1,1}(y,X\gamma_d)K^{(1)}((X\gamma_d - x\gamma_d)/h)\right] s_{d,1}(dy,x\gamma_d)\\ =& h^{-1}\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\cdot\left\{\int_{\mathbb{R}} K^{(1)}(u)\rho^{\tilde{\gamma}_d}_{1,1}(y,x\tilde{\gamma}_d+uh)f_d(x\tilde{\gamma}_d+uh)du\right.\\ &\left.-\int_{\mathbb{R}} K^{(1)}(u)\rho^{\gamma_d}_{1,1}(y,x\gamma_d+uh)f_d(x\gamma_d+uh)du\right\}s_{d,1}(dy,x\gamma_d)\\ =& \int_{\mathbb{R}} uK^{(1)}(u)du\cdot\int_0^t \ddot{\phi}^{\theta}_{d,\gamma_d}(y,x)\cdot\left\{\left(\partial z\rho^{\tilde{\gamma}_d}_{1,1}(y,z)|_{z=x\tilde{\gamma}_d}f_d(x\tilde{\gamma}_d) + \rho^{\tilde{\gamma}_d}_{1,1}(y,x\tilde{\gamma}_d)\partial_z f_d(z)|_{z=x\tilde{\gamma}_d}\right)\right.\\ &\left.-\left(\partial z\rho^{\gamma_d}_{1,1}(y,z)|_{z=x\gamma_d}f_d(x\gamma_d) + \rho^{\gamma_d}_{1,1}(y,x\gamma_d)\partial_z f_d(z)|_{z=x\gamma_d}\right)\right\}s_{d,1}(dy,x\gamma_d),\end{aligned}$$

where $\rho^{\gamma}_{1,1}$ is defined in (1.4.8). The second equality follows by Taylor expansion and the fact that $\int_{[-1,1]} K^{(1)}(u)du = 0$. By the Lipschitz continuity of $\rho^{\gamma}_{1,1}(y,x\gamma), \partial_{x\gamma}\rho^{\gamma}_{1,1}(y,x\gamma), f_d(x\gamma)$, and $\partial_{x\gamma}f_d(x\gamma)$, with respect to γ as implied by Assumption 1.7.1, and by the fact that $\|\tilde{\gamma}_d - \gamma_d\| \leq \delta_n$, we conclude that $\sup_{\|\tilde{\gamma}_d-\gamma_d\|\leq\delta_n}\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\|\Delta_{11}(t,x,\theta,\tilde{\gamma}_d)\| = O(\delta_n)$.

The centered term Δ_{12} can be bounded using following empirical process

$$\mathscr{G}_{\delta,n} \equiv \{w \mapsto g_\delta(w,\omega,\tilde{\gamma}_d) : \omega \in \Omega, \|\tilde{\gamma}_d - \gamma_d\| \leq \delta_n\},$$

where $g_\delta(W,\omega,\tilde{\gamma}_d) \equiv g_{\delta,1}(W,\omega,\tilde{\gamma}_d) - \int g_{\delta,1}(W,\omega,\tilde{\gamma}_d)dF_W(W)$, and

$$\begin{aligned} g_{\delta,1}(W,\omega,\tilde{\gamma}_d) \equiv & \int_0^t \ddot{\phi}^\theta_{d,\gamma_d}(y,x)\mathbb{1}\{D=d,Y\leq y\}s_{d,1}(dy,x\gamma_d) \\ & \cdot \left\{K^{(1)}((X\tilde{\gamma}_d - x\tilde{\gamma}_d)/h) - K^{(1)}((X\gamma_d - x\gamma_d)/h)\right\}. \end{aligned}$$

Applying similar lines of arguments as in Lemma 1.7, it is straightforward to show that $\mathscr{G}_{\delta,n}$ is a VC type class with bounded envelop, for each δ_n. From the continuous differentiability of $K^{(1)}(\cdot)$, we deduce by similar arguments to Lemma 8.4 in Maistre and Patilea (2019) that $\left|K^{(1)}((X\tilde{\gamma}_d - x\tilde{\gamma}_d)/h) - K^{(1)}((X\gamma_d - x\gamma_d)/h)\right| \leq \delta_n h^{-1}||K^{(2)}((X\gamma_d - x\gamma_d)/h)\ || + C\delta_n^2 h^{-2}$, for some positive constant C. Combine this fact with the uniform boundedness of $\ddot{\phi}^\theta_{d,\gamma_d}$, and $s_{d,1}$, and we find that $\sup_{g_\delta \in \mathscr{G}_\delta} \mathbb{E}[g_\delta^2]$ is bounded from above at the rate of $O\left(\delta_n^2 h^{-1}\right)$. We then conclude from applying the maximal inequality in Lemma 1.5 that $\sup_{\|\tilde{\gamma}_d - \gamma_d\|} \sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta} \|\Delta_{12}(t,x,\theta,\tilde{\gamma}_d)\| = O_p\left((\log n)^{1/2} n^{-1/2} h^{-5/2}\delta_n\right)$.

For Δ_2, we have

$$\begin{aligned} \Delta_2(t,x,\theta,\tilde{\gamma}_d) = & - f_d^{-2}(x\tilde{\gamma}_d)\int_0^t \ddot{\phi}^\theta_{d,\gamma_d}(y,x)\left\{\hat{\kappa}_{d,y}(x\tilde{\gamma}_d) - \hat{\kappa}_{d,y}(x\gamma_d)\right\}\partial_\gamma \hat{f}_d(x\tilde{\gamma}_d)s_{d,1}(dy,x\gamma_d) \\ & - f_d^{-2}(x\gamma_d)\int_0^t \ddot{\phi}^\theta_{d,\gamma_d}(y,x)\kappa_{d,y}(x\gamma_d)\left\{\partial_\gamma \hat{f}_d(x\tilde{\gamma}_d) - \partial_\gamma \hat{f}_d(x\gamma_d)\right\}s_{d,1}(dy,x\gamma_d) \\ & + \frac{(f_d(x\tilde{\gamma}_d) + f_d(x\gamma_d))(\hat{f}_d(x\tilde{\gamma}_d) - \hat{f}_d(x\gamma_d))}{f_d^{-2}(x\tilde{\gamma}_d)f_d^{-2}(x\gamma_d)} \\ & \cdot \int_0^t \ddot{\phi}^\theta_{d,\gamma_d}(y,x)\kappa_{d,y}(x\gamma_d)\partial_\gamma \hat{f}_d(x\gamma_d)s_{d,1}(dy,x\gamma_d). \end{aligned}$$

Arguing as in the case of Δ_1, one finds that the second term in the above display dominates the other two with a uniform rate of $O_p\left((\log n)^{1/2} n^{-1/2} h^{-5/2}\delta_n\right) + O(\delta_n)$. Gathering results on Δ_1 and Δ_2 completes our proof. ■

Lemma 1.9 Suppose the conditions of Theorem 1.5 hold. Then

$$\begin{aligned} \sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta} & \left| n^{-1/2}h^{1/2}\sum_{i=1}^n g_{d,\gamma_d,\ell}(X_i,x)\left\{\hat{\Psi}_d\left(\hat{\mathscr{E}}_{d,\hat{\gamma}_d,i}, \hat{\mathscr{E}}_{d,1,\hat{\gamma}_d,i}\right)(t,x,\theta) - \Psi_d\left(\mathscr{E}_{d,\gamma_d,i}, \mathscr{E}_{d,1,\gamma_d,i}\right)(t,x,\theta)\right\}\right| \\ & = O_p\left((\log n)^{1/2} n^{-1/2} h^{-(2\ell+1)/2}\right), \end{aligned}$$

where $g_{d,\gamma_d,\ell}(X,x) = h^{-(\ell+1)}K^{(\ell)}(x\gamma_d, X\gamma_d)/f(x\gamma_d,d)$, for $\ell = 0,1$.

Proof of Lemma 1.9. Define $\eta_{3,1}(t,x\gamma) \equiv \phi_\theta'\left(s_{T_d}(t,x\gamma,\theta)\right)$, $\eta_{3,2}(t,x\gamma) \equiv \phi_\theta''\left(s_d(t,x\gamma)\right)$, $\eta_{3,3}(W,t,\gamma) \equiv \mathbb{1}\{D=d\}\cdot (\mathbb{1}\{Y\leq t\} - G_d(t,X\gamma))$, $\eta_{3,4}(t,x\gamma) \equiv s_{d,1}(t,x\gamma)$, $\eta_{3,5}(W,t,\gamma) \equiv \mathbb{1}\{D=d\}\left(R\mathbb{1}\{Y\leq t\} - G_{d,1}(t,X\gamma)\right)$, and for $\ell = 1,...,5$, let the estimator of $\eta_{3,\ell}$ be denoted by $\hat{\eta}_{3,\ell}$. Their definitions should be apparent. Index on θ is suppressed.

From Theorem 1.3 and Lemma 1.4, we have

$$\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta} \left|\hat{\eta}_{3,\ell}(t,x\gamma_d) - \eta_{3,\ell}(t,x\gamma_d)\right| = O_p\left((\log n)^{1/2} n^{-1/2} h^{-1/2}\right),$$

for $\ell = 1,2,4$.

Given these notations, we divide Ψ_d into

$$\begin{aligned}
\Psi_{d,1}(W,t,x\gamma) &= \frac{1}{\eta_{3,1}(t,x\gamma)}\int_0^t \eta_{3,2}(y,x\gamma)\eta_{3,3}(W,y,x\gamma)\eta_{3,4}(dy,x\gamma),\\
\Psi_{d,2}(W,t,x\gamma) &= \frac{-1}{\eta_{3,1}(t,x\gamma)}\eta_{3,4}(t,x\gamma)\eta_{3,5}(W,t,x\gamma),\\
\Psi_{d,3}(W,t,x\gamma) &= \frac{1}{\eta_{3,1}(t,x\gamma)}\int_0^t \eta_{3,2}(y,x\gamma)\eta_{3,5}(W,y,x\gamma)\eta_{3,4}(dy,x\gamma),
\end{aligned}$$

and thus $\Psi_d(\mathscr{E}_{d,\gamma},\mathscr{E}_{d,1,\gamma}) = \sum_{\ell=1}^3 \Psi_{d,\ell}$. We illustrate on $\Psi_{d,1}$, since the other two terms share a similar structure. From tedious manipulation, it can be shown that $\hat{\Psi}_{d,1}(W,t,x\hat{\gamma}_d) - \Psi_{d,1}(W,t,x\gamma_d) = \sum_{\ell=1}^{10} A_{3,\ell}(W,t,x)$, where

$$\begin{aligned}
A_{3,1}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)}\int_0^t \eta_{3,2}(y,x\gamma_d)\eta_{3,3}(W,y,\gamma_d)\eta_{3,4}(dy,x\gamma_d),\\
A_{3,2}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)}\int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d))\eta_{3,3}(W,y,\gamma_d)\eta_{3,4}(dy,x\gamma_d),\\
A_{3,3}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)}\int_0^t \eta_{3,2}(y,x\gamma_d)(\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))\eta_{3,4}(dy,x\gamma_d),\\
A_{3,4}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)}\int_0^t \eta_{3,2}(y,x\gamma_d)\eta_{3,3}(W,y,\gamma_d)(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)),\\
A_{3,5}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)}\\
&\cdot\int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d))\eta_{3,3}(W,y,\gamma_d)\eta_{3,4}(dy,x\gamma_d),\\
A_{3,6}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)}\\
&\cdot\int_0^t \eta_{3,2}(y,x\gamma_d)(\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))\eta_{3,4}(dy,x\gamma_d),\\
A_{3,7}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)}\\
&\cdot\int_0^t \eta_{3,2}(y,x\gamma_d)\eta_{3,3}(W,y,\gamma_d)(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)),
\end{aligned}$$

$$
\begin{aligned}
A_{3,8}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)} \int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d)) \\
& \cdot (\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))\eta_{3,4}(dy,x\gamma_d), \\
A_{3,9}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)} \int_0^t \eta_{3,2}(y,x\gamma_d)(\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d)) \\
& \cdot (\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)), \\
A_{3,10}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)} \int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d)) \\
& \cdot \eta_{3,3}(y,x\gamma_d)(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)), \\
A_{3,11}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)} \\
& \cdot \int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d))(\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))\eta_{3,4}(dy,x\gamma_d), \\
A_{3,12}(W,t,x) =& \frac{1}{\eta_{3,1}(t,x\gamma_d)} \int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d)) \\
& \cdot (\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)), \\
A_{3,13}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)} \\
& \cdot \int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d))\eta_{3,3}(W,y,\gamma_d)(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)), \\
A_{3,14}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)} \\
& \cdot \int_0^t \eta_{3,2}(y,x\gamma_d)(\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)), \\
A_{3,15}(W,t,x) =& -\frac{\hat{\eta}_{3,1}(t,x\hat{\gamma}_d) - \eta_{3,1}(t,x\gamma_d)}{\eta_{3,1}(t,x\gamma_d)\hat{\eta}_{3,1}(t,x\hat{\gamma}_d)} \cdot \int_0^t (\hat{\eta}_{3,2}(y,x\hat{\gamma}_d) - \eta_{3,2}(y,x\gamma_d)) \\
& \cdot (\hat{\eta}_{3,3}(W,y,\hat{\gamma}_d) - \eta_{3,3}(W,y,\gamma_d))(\hat{\eta}_{3,4}(dy,x\hat{\gamma}_d) - \eta_{3,4}(dy,x\gamma_d)).
\end{aligned}
$$

Following the same type of analysis we have used so far, namely performing Taylor expansion, integration by parts, and applying the maximal inequality from Lemma 1.5 whenever appropriate, we get

$$
\begin{aligned}
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta} \left| n^{-1/2}h^{1/2}\sum_{i=1}^n g_{d,\gamma_d,\ell}(X_i,x)A_{3,\ell_1}(W_i,t,x) \right| &= O_p\left((\log n)^{1/2} n^{-1/2} h^{-(2\ell+1)/2} \right), \\
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta} \left| n^{-1/2}h^{1/2}\sum_{i=1}^n g_{d,\gamma_d,\ell}(X_i,x)A_{3,\ell_2}(W_i,t,x) \right| &= O_p\left(\log n \cdot n^{-1} h^{-(\ell+1)} \right), \\
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta} \left| n^{-1/2}h^{1/2}\sum_{i=1}^n g_{d,\gamma_d,\ell}(X_i,x)A_{3,\ell_3}(W_i,t,x) \right| &= O_p\left((\log n)^{3/2} n^{-3/2} h^{-(2\ell+3)/2} \right), \\
\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta} \left| n^{-1/2}h^{1/2}\sum_{i=1}^n g_{d,\gamma_d,\ell}(X_i,x)A_{3,15}(W_i,t,x) \right| &= O_p\left((\log n)^2 n^{-2} h^{-(\ell+2)} \right).
\end{aligned}
$$

for $\ell = 0, 1$, $\ell_1 = 1, 2, 3, 4$, $\ell_2 = 5, 6, 7, 8, 9, 10$, and $\ell_3 = 11, 12, 13, 14$. As a result,

$$\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|n^{-1/2}h^{1/2}\sum_{i=1}^{n} g_{d,\gamma_d,\ell}(X_i,x)\left(\hat{\Psi}_{d,1}(W_i,t,x\hat{\gamma}_d)-\Psi_{d,1}(W_i,t,x\gamma_d)\right)\right| = O_p\left((\log n)^{1/2}n^{-1/2}h^{-(2\ell+1)/2}\right).$$

Analogous results hold for $\Psi_{d,2}$ and $\Psi_{d,3}$, concluding the proof. ∎

1.9.3.3 Covariance Functions

Lemma 1.10 Suppose the assumptions of Corollary 1.1 hold. Then, it holds that, $\Sigma_\eta^x(\cdot,\cdot) = \Sigma_\eta^{x\dagger}(\cdot,\cdot) + o(1)$, where

$$\Sigma_\eta^{x\dagger}(\mathbf{t},\boldsymbol{\theta}) = \begin{pmatrix} \sigma_{1,x}^2(t_1,t_2,\theta_1,\theta_2) & 0 \\ 0 & \sigma_{0,x}^2(t_1,t_2,\theta_3,\theta_4) \end{pmatrix},$$

and

$$\begin{aligned}
&\sigma_{d,x}^2(t_1,t_2,\theta_1,\theta_2) = \frac{\|K\|_2^2}{f_d(x\gamma_d)\phi'_{\theta_1}(s_{T_d}(t_1,x\gamma_d,\theta_1))\phi'_{\theta_2}(s_{T_d}(t_2,x\gamma_d,\theta_2))} \\
&\cdot\left\{\int_0^{t_1}\int_0^{t_2}\ddot{\phi}_{d,\gamma_d}^{\theta_1}(y_1,x)\ddot{\phi}_{d,\gamma_d}^{\theta_2}(y_2,x)\right. \\
&\qquad\cdot\{G_d(y_1\wedge y_2,x\gamma_d)-G_d(y_1,x\gamma_d)G_d(y_2,x\gamma_d)\}\,s_{d,1}(dy_2,x\gamma_d)s_{d,1}(dy_1,x\gamma_d) \\
&+\int_0^{t_1}\ddot{\phi}_{d,\gamma_d}^{\theta_1}(y_1,x)\dot{\phi}_{d,\gamma_d}^{\theta_2}(t_2,x)\left\{G_{d,1}(y_1\wedge t_2,x\gamma_d)-G_d(y_1,x\gamma_d)G_{d,1}(t_2,x\gamma_d)\right\}s_{d,1}(dy_1,x\gamma_d) \\
&-\int_0^{t_1}\int_0^{t_2}\ddot{\phi}_{d,\gamma_d}^{\theta_1}(y_1,x)\ddot{\phi}_{d,\gamma_d}^{\theta_2}(y_2,x) \\
&\qquad\cdot\left\{G_{d,1}(y_1\wedge y_2,x\gamma_d)-G_d(y_1,x\gamma_d)G_{d,1}(y_2,x\gamma_d)\right\}s_d(dy_2,x\gamma_d)s_{d,1}(dy_1,x\gamma_d) \\
&+\int_0^{t_2}\dot{\phi}_{d,\gamma_d}^{\theta_1}(t_1,x)\ddot{\phi}_{d,\gamma_d}^{\theta_2}(y_2,x)\left\{G_{d,1}(y_2\wedge t_1,x\gamma_d)-G_d(y_2,x\gamma_d)G_{d,1}(t_1,x\gamma_d)\right\}s_{d,1}(dy_2,x\gamma_d) \\
&+\dot{\phi}_{d,\gamma_d}^{\theta_1}(t_1,x)\dot{\phi}_{d,\gamma_d}^{\theta_2}(t_2,x)\left\{G_{d,1}(t_1\wedge t_2,x\gamma_d)-G_{d,1}(t_1,x\gamma_d)G_{d,1}(t_2,x\gamma_d)\right\} \\
&-\int_0^{t_2}\dot{\phi}_{d,\gamma_d}^{\theta_1}(t_1,x)\ddot{\phi}_{d,\gamma_d}^{\theta_2}(y_2,x)\left\{G_{d,1}(t_1\wedge y_2,x\gamma_d)-G_{d,1}(y_2,x\gamma_d)G_{d,1}(t_1,x\gamma_d)\right\}s_d(dy_2,x\gamma_d) \\
&-\int_0^{t_1}\int_0^{t_2}\ddot{\phi}_{d,\gamma_d}^{\theta_1}(y_1,x)\ddot{\phi}_{d,\gamma_d}^{\theta_2}(y_2,x) \\
&\qquad\cdot\left\{G_{d,1}(y_1\wedge y_2,x\gamma_d)-G_d(y_2,x\gamma_d)G_{d,1}(y_1,x\gamma_d)\right\}s_{d,1}(dy_2,x\gamma_d)s_d(dy_1,x\gamma_d) \\
&-\int_0^{t_1}\dot{\phi}_{d,\gamma_d}^{\theta_2}(t_2,x)\ddot{\phi}_{d,\gamma_d}^{\theta_1}(y_1,x)\left\{G_{d,1}(y_1\wedge t_2,x\gamma_d)-G_{d,1}(t_2,x\gamma_d)G_{d,1}(y_1,x\gamma_d)\right\}s_d(dy_1,x\gamma_d) \\
&+\int_0^{t_1}\int_0^{t_2}\ddot{\phi}_{d,\gamma_d}^{\theta_1}(y_1,x)\ddot{\phi}_{d,\gamma_d}^{\theta_2}(y_2,x) \\
&\qquad\left.\cdot\left\{G_{d,1}(y_1\wedge y_2,x\gamma_d)-G_{d,1}(y_1,x\gamma_d)G_{d,1}(y_2,x\gamma_d)\right\}s_d(dy_2,x\gamma_d)s_d(dy_1,x\gamma_d)\right\}
\end{aligned} \tag{1.9.38}$$

Proof of Lemma 1.10. When $d_1 \neq d_2$, $\Psi_{d_1}(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}) \cdot \Psi_{d_2}(\mathscr{E}_{d,\gamma_d},\mathscr{E}_{d,1,\gamma_d}) = 0$, implying the off-diagonal element of the covariance matrix is 0, regardless of t and θ. For terms on the main diagonal, the proof is analogous to that of Lemma C.1 in Fan and Liu (2018), and thus, we omit the details. ■

1.9.3.4 First-Stage Estimator for the QTE

To estimate the bound curves for the QTE, we replace $f_{T_d,x}(t,\theta)$ and $f_{T_d}(t,\theta)$, with preliminary estimates $\hat{f}_{T_d,x}(t,\theta)$ and $\hat{f}_{T_d}(t,\theta)$, respectively. Validity of multiplier bootstrap procedure in Section 1.5.1 hinges on the estimates being uniformly consistent in (t,θ). In what follows, we provide estimators, based on the analytical expression of f_{T_d}, that satisfy this property. Using the closed form expression for s_{T_d} from Theorem 1.2, we deduce that

$$f_{T_d,x}(t,\theta) = \frac{\phi'_\theta(s_d(t,x\gamma_d))}{\phi'_\theta(s_{T_d}(t,x\gamma_d,\theta))} \cdot f_{d,1}(t,x\gamma_d),$$

where $f_{d,1}(t,x\gamma) = -\partial_t s_{d,1}(t,x\gamma)$. The fraction in the above display can be estimated by reusing $\hat{\gamma}_d$, $\hat{s}_d$, and $\hat{s}_{T_d}$ from (1.4.2), (1.4.4), and (1.4.5), respectively. For $f_{d,1}(t,x\gamma)$, we will use the SIM conditional density estimator as follows.

Let $H^x(\cdot)$ and $H^y(\cdot)$ be kernel functions associated with Y and $X\gamma$, respectively. Given the sequences of bandwidth λ_x and λ_y that fulfill the conditions in Assumption 1.13, the rescaled kernel are once again defined by $H^j_\lambda(u,v) = \lambda^{-1}H(\lambda^{-1}(v-u))$, for $j = x,y$. Now we let

$$\hat{f}_{d,1}(t,x\gamma) = \frac{\sum_{i=1}^n \mathbb{1}\{D_i = d, R_i = 1\} \cdot H^y_{\lambda_y}(t,Y_i) \cdot H^x_{\lambda_x}(x\gamma, X_i\gamma)}{\sum_{i=1}^n \mathbb{1}\{D_i = d\} \cdot H^x_{\lambda_x}(x\gamma, X_i\gamma)}, \tag{1.9.39}$$

In addition, we will have $f_{T_d,x}$ estimated by $\hat{f}_{T_d,x}(t,\theta) = \phi'_\theta(\hat{s}_d(t,x\hat{\gamma}_d))\hat{f}_{d,1}(t,x\hat{\gamma}_d)/\phi'_\theta(\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta))$. For the unconditional density f_{T_d}, we can estimate it by taking the sample average of $\hat{f}_{T_d,X}(t,\theta)$ with respect to X, i.e. $\hat{f}_{T_d}(t,\theta) \equiv n^{-1}\sum_{i=1}^n \hat{f}_{T_d,X_i}(t,\theta)$.

Assumption 1.13 For $j = x,y$, (i) the kernel function, $H^j(\cdot)$ is symmetric, compactly supported, and of bounded variation;[13] (ii) it is twice continuously differentiable and the second order derivative is continuous and of bounded variation; (iii) $\lambda_j \to 0$, $\log n \cdot n^{-1/2}\lambda_y\lambda_x \to 0$, as $n \to \infty$. For $d \in \{0,1\}$, (iv) $v \mapsto \rho^\gamma_{f,d}(y,v)$, where $\rho^\gamma_{f,d}(y,x\gamma) \equiv f_d(x\gamma)\mathbb{E}\left[\partial_y G_{d,1}(y,X\gamma_d)|X\gamma = x\gamma\right]$, is continuously differentiable and the derivative is bounded uniformly on $\tilde{\mathscr{T}} \times \mathscr{X} \times \Gamma_{d,n}$; (v) $\partial_v\rho^\gamma_{f,d}(y,v)$ is Lipschitz continuous in v with the Lipschitz constant being independent of y, x, and $\gamma \in \Gamma_{d,n}$.

[13] The compactness assumption could be relaxed at the expanse of longer proof, and therefore, the Gaussian kernel could be accommodated.

Lemma 1.11 Under the assumptions of Corollary 1.2 and Assumption 1.13, it holds that $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}|\hat{f}_{T_d,x}(t,\theta)$ $-f_{T_d,x}(t,\theta)| = o_p(1)$, and $\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|\hat{f}_{T_d}(t,\theta)-f_{T_d}(t,\theta)\right| = o_p(1)$, for $d\in\{0,1\}$.

Proof of Lemma 1.11. We first show that $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|\hat{f}_{d,1}(t,x\hat{\gamma}_d)-f_{d,1}(t,x\gamma_d)\right| = o_p(1)$. By the triangular inequality, we have that

$$\begin{aligned}\left|\hat{f}_{d,1}(t,x\hat{\gamma}_d)-f_{d,1}(t,x\gamma_d)\right| \leq& \left|\hat{f}_{d,1}(t,x\hat{\gamma}_d)-\hat{f}_{d,1}(t,x\gamma_d)\right|+\left|\hat{f}_{d,1}(t,x\gamma_d)-f_{d,1}(t,x\gamma_d)\right| \\ \equiv& \Delta_{f,1}(t,x,\hat{\gamma}_d)+\Delta_{f,2}(t,x,\hat{\gamma}_d).\end{aligned} \tag{1.9.40}$$

Along lines of arguments similar to that of Lemma 1.8, we get that $\sup_{\|\gamma-\gamma_d\|\leq\delta_n}\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}\left|\Delta_{f,1}(t,x,\gamma,\theta)\right| = O_p\left((\log n)^{1/2}n^{-1/2}\lambda^{-3/2}\delta_n\right)+O(\delta_n)$. The second part arises from bias calculations, which depends crucially on Assumptions 1.13(iv) and 1.13(v). Next, it follows, by standard bias calculation and direct application of Theorems 1 and 4 in Einmahl and Mason (2005), that $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|\Delta_{f,2}(t,x)\right| = O_p\left((\log n)^{1/2}n^{-1/2}\lambda_y^{-1/2}\lambda_x^{-1/2}+\lambda_x^2+\lambda_y^2\right)$, which is $o_p(1)$ under Assumption 1.13(iii). Combining these results, we conclude that the left hand side of (1.9.40) is $o_p(1)$.

Now, observe that

$$\begin{aligned}\hat{f}_{T_d,x}(t,\theta)-f_{T_d,x}(t,\theta) =& \left\{\frac{\phi_\theta^{'}(\hat{s}_d(t,x\hat{\gamma}_d))}{\phi_\theta^{'}(\hat{s}_{T_d}(t,x\hat{\gamma}_d,\theta))}-\frac{\phi_\theta^{'}(s_d(t,x\gamma_d))}{\phi_\theta^{'}(s_{T_d}(t,x\gamma_d,\theta))}\right\}\hat{f}_{d,1}(t,x\hat{\gamma}_d) \\ &+\frac{\phi_\theta^{'}(s_d(t,x\gamma_d))}{\phi_\theta^{'}(s_{T_d}(t,x\gamma_d,\theta))}\left\{\hat{f}_{d,1}(t,x\hat{\gamma}_d)-f_{d,1}(t,x\gamma_d)\right\}.\end{aligned}$$

From the fact that $\hat{s}_d$ and $\hat{s}_{T_d}$ are uniformly convergent, that $\dot{\phi}_\theta^{-1}(z)$ is uniformly bounded away from 0 on $[0,y_o^*]$, and that, for each $(t,x)\in\tilde{\mathscr{T}}\times\mathscr{X}$, $\phi(\hat{s}_{T_d}(t,x\hat{\gamma}_d))$ belongs to $[0,y_o^*]$ with probability approaching 1, we deduce that the difference inside the curly braces in the first line is $o_p(1)$. Under Assumption 1.6.3, we have $\phi_\theta^{'}(s_d(t,x\gamma_d))/\phi_\theta^{'}(s_{T_d}(t,x\gamma_d,\theta))$ is uniformly $O(1)$.

The function $f_{d,1}(\cdot,\cdot\gamma_d)$ is uniformly bounded from Assumption 1.6.2. It then follows from the uniform convergence results we derived earlier, that $\hat{f}_{d,1}(t,x\hat{\gamma}_d)$ is also uniformly bounded across $(t,x)\in\tilde{\mathscr{T}}\times\mathscr{X}$. Consequently, $\sup_{(t,x,\theta)\in\tilde{\mathscr{T}}\times\mathscr{X}\times\Theta}\left|\hat{f}_{T_d,x}(t,\theta)-f_{T_d,x}(t,\theta)\right| = o_p(1)$. We notice that this also implies that $\sup_{(t,\theta)\in\tilde{\mathscr{T}}\times\Theta}|\hat{f}_{T_d}(t,\theta)-$ $f_{T_d}(t,\theta)| = o_p(1)$, for $d\in\{0,1\}$, which concludes our proof. ■

CHAPTER 2

Two Sample Unconditional Quantile Effect

This chapter is adapted from the working paper "Two Sample Unconditional Quantile Effect" and has been reproduced with the permission of my co-authors Atsushi Inoue and Tong Li.

2.1 Introduction

Missing data is a ubiquitous problem in empirical studies. Consider the scenario where a researcher is interested in conducting counterfactual analysis on a target variable, but it is entirely missing from the dataset of interest. In such circumstances, counterfactual policy effects cannot be identified from the primary dataset alone, and therefore, external information and/or stronger identifying assumptions are necessary. In this paper, we utilize both to achieve identification. Specifically, we focus on the situation where the missing variable can be found in another dataset and the information from which can be used to recover target policy parameters in the population of interest, under a set of commonly assumed restrictions on both the data structure and the model primitives.

To fix ideas, consider the following example. Suppose we are interested in studying the effect of a counterfactual change in the distribution of actual labor market experience on some distributional features of yearly earnings. Our main dataset does not record respondents' work history, and therefore, we cannot recover their actual labor market experience. Suppose the variable is available from a second dataset, but it may not be a reliable source of information on income or it may not be representative of the target population we aim to analyze. In this case, we would benefit from combining information from both samples to identify and estimate our parameter of interest.

Research on counterfactual policy effects under data combination is scarce. Our paper fills this gap by proposing a new framework that accommodates such a data structure. In this paper, we focus on one particular type of counterfactual policy effects, the *unconditional quantile effect* (UQE). It measures the effect of a marginal change in the unconditional distribution of a single covariate on the quantiles of a target outcome. We provide identification results for UQE under various types of marginal distributional change. The key insight of our identification strategy is that some covariates present in both datasets can be excluded from the outcome equation, which would provide a source of exogenous variations that allows us to recover the joint distribution of missing variables, otherwise not identified using the two samples separately.

The second contribution of the paper is to propose novel semiparametric estimators based on these identification results. Departing from the literature on the estimation of counterfactual quantile effects—see, e.g. Firpo et al.

(2009b), Sasaki et al. (2022), etc.—which focuses primarily on the *marginal location shift* (MLS) of a covariate, we provide estimators of UQE under two general types of counterfactual distributional changes, namely the *marginal distributional shift* (MDS) and the *marginal quantile shift* (MQS),[1] the latter of which includes MLS as a special case. To the best of our knowledge, large sample results for these two cases are new to the literature. We apply these results to study a variant of Mincer's earnings function. Using data from Integrated Public Use Microdata Sample (IPUMS) as our main data source and the Panel Study of Income Dynamics (PSID) as the auxiliary sample, we investigate the counterfactual effect of actual work experiences on income. The effect profiles with MDS and MQS are found to be similar in shape.

This paper belongs to the growing literature on the (marginal) unconditional policy effect. Since Firpo et al. (2009b) introduced the method of *unconditional quantile regressions* (UQR), the study of unconditional policy effect has gained much attention. In general, this parameter differs from the one identified by the conditional quantile regression (Koenker and Bassett Jr, 1978), where marginal effects on the conditional quantile are the locus of attention. Applied researcher are often interested in the shifts in the quantiles of unconditional distribution of a target outcome. For instance, one may take an interest in how wage distribution changes in response to marginal increases in some characteristics of the labor force, such as education level and experience. Conditional quantile regression cannot be applied to address this type of questions, whereas UQR suits the goal.

Rothe (2012) generalizes the method of Firpo et al. (2009b), and analyzes a variety of counterfactual policy effects. He formalizes the idea of *ceteris paribus* distributional change and provides extensive results for both fixed and marginal policy shifts. Our identification framework is closely related to his treatment of the latter type. Focusing on the special case of quantile effects, we extend his identification results to a data combination setting and provide novel inference theories specifically tailored to the distinct features of combined samples. For recent development in this literature, see Firpo et al. (2018), Martinez-Iriarte and Sun (2020), Martínez-Iriarte (2023), and Sasaki et al. (2022). For a comprehensive survey on counterfactual distributions and decomposition methods, see Fortin et al. (2011).

Our paper also builds on the econometric methods of data combination. In economics, this strand of literature stems from the *two-sample instrumental variables* (TSIV) model that was first introduced by Klevmarken (1982), Angrist and Krueger (1992), Arellano and Meghir (1992), and is later extended by Ridder and Moffitt (2007), Inoue and Solon (2010), among others. Conceptually, the semiparametric data combination model we consider here is different from the traditional missing data problem (Robins et al., 1994). It is more closely related to the "verify-out-of-sample" model in Chen et al. (2008), and also to Imbens and Lancaster (1994), Fan et al. (2014), Graham et al. (2016), Hirukawa et al. (2020), and Buchinsky et al. (2022), to name a few.

[1] The precise definitions of MLS, MDS, and MQS are given in Section 2.3.

The paper is organized as follows. In the next section, we describe the model and assumptions on the data structure. In Section 2.3 we introduce the parameter of interest, and then present identification results for continuously distributed and discrete target covariates, respectively. Section 2.4 discusses the estimation strategy and large sample results. We apply the method to study the income effect of real labor market experience in Section 2.5. Section 2.6 concludes. Proofs and auxiliary results are collected in Sections 2.7 and 2.8.

2.2 Setup

The objective of our paper is to analyze the effect of a counterfactual change in the marginal distribution of the covariate of interest, X, on the quantiles of the target outcome, Y, under data combination. The precise definition of the counterfactual policy effect is provided in Section 2.3. When X is exogenous, and all the variables relevant for analysis are observed from a single data source, counterfactual policy effects can be analyzed either directly by applying tools from Firpo et al. (2009b) and Rothe (2012), or indirectly by recovering the structural function using standard identification results such as Matzkin (2003) and Matzkin (2007). However, when the variables of interest are scattered among several different data sources, we face a fundamental identification problem: The conditional distribution of Y given X is not identified from any single sample. In this case, existing methods do not provide an immediate solution.

Throughout this paper, we consider the scenario where our Y and X are sourced from two different data sets. The outcome is contained in the *principal* or *main* sample, $\mathscr{S}_s = \{Y_i, Z_i\}_{i=1}^{n_s}$, from the *study* population, $\mathscr{P}_s$. The target covariate is missing completely from $\mathscr{S}_s$. However, it is observed in the *auxiliary* sample, $\mathscr{S}_a = \{X_i, Z_i\}_{i=1}^{n_a}$, from the *auxiliary* population, $\mathscr{P}_a$, which does not contain observations of Y.

We now formally describe our structural model. We allow variables from two populations to be determined by different mechanisms. For the study population,

$$Y_s = g_s(X_s, Z_1, \varepsilon_s), \tag{2.2.1}$$

$$X_s = h_s(Z, \eta_s), \tag{2.2.2}$$

where $Y_s \in \mathscr{Y} \subset \mathbb{R}$ is the potential outcome in the study population, $\varepsilon_s \in \mathscr{E} \subset \mathbb{R}^{d_\varepsilon}$ is a vector of unobserved heterogeneity term. Equation (2.2.1), links the target outcome, a scalar variable, $X_s \in \mathscr{X} \subset \mathbb{R}$, and a vector of exogenous variables, Z_1. Here, X_s is the potential covariate of interest in the study population, which is in turn determined by (2.2.2). We can think of (2.2.2) as the reduced form relationship between X_s and Z, where $Z = (Z_1', Z_2')' \in \mathscr{Z} \equiv \mathscr{Z}_1 \times \mathscr{Z}_2 \subset \mathbb{R}^{d_z}$ includes both the exogenous variables in the outcome equation and a vector of excluded instrument, Z_2. The vector of instrument, Z, is available in both samples, and therefore, it serves to establish a link between two samples. The model in (2.2.1) accommodates general nonseparability between covariates and the unobserved heterogeneity. We do not impose any parametric or shape restriction on g_s.

Variables in the auxiliary population are determined by

$$Y_a = g_a(X_a, Z_1, \varepsilon_a) \qquad \text{and} \qquad X_a = h_a(Z, \eta_a),$$

where g_a and h_a are generally different from g_s and h_s, respectively.

Let R denote the sample membership indicator. That is, $R_i = 1$, if i-th draw comes from the study population, $i = 1, ..., n = n_s + n_a$. Let $Y = RY_s + (1-R)Y_a$ and $X = RX_s + (1-R)X_a$. If no variable is missing, we are able to observe (Y_s, Y_a, X_s, X_a). However, in our context, only RY and $(1-R)X$ are observed. We then construct a pseudo-merged sample $\mathscr{S}$ using the two data sources as $\mathscr{S} = \{R_i, R_iY_i, (1-R_i)X_i, Z_i\}_{i=1}^n$. Let $A = (R, RY, (1-R)X, Z)$ and $W = (X', Z_1')'$ collect the observed variables and the covariates in the outcome equation, respectively. Throughout the paper, we arrange the data in such a way that $R_i = 1$ for $i = 1, ..., n_s$ and $R_i = 0$ for $i = n_s + 1, ..., n$. The merged sample may not correspond to any real-world population. We impose the following set of assumptions on the merged sample so it can mimic a random sample from a pseudo population. These assumptions are largely based on Assumption 1 in Graham et al. (2016).

Assumption 2.1 (Data Structure)

(a) $\text{Supp}(F_{Z|R=1}) \subset \text{Supp}(F_{Z|R=0})$.

(b) (i) $n_s/(n_s + n_a) \to Q_0$; (ii) R follows a Bernoulli distribution, with $\mathbb{E}[R] = Q_0$.

(c) There is a unique measurable function $r(\cdot) : \mathscr{Z} \to [0,1]$, such that for all $z \in \mathscr{Z}$,

$$\frac{f_{Z|R}(z|1)}{f_{Z|R}(z|0)} = \frac{1-Q_0}{Q_0} \frac{r(z)}{1-r(z)}.$$

(d) (i) $Q_0 \in (\varepsilon_1, 1-\varepsilon_1)$ for some $\varepsilon_1 \in (0, 1/2)$; (ii) $\varepsilon_2 < r(z) < 1 - \varepsilon_2$ for some $\varepsilon_2 \in (0, 1/2)$, and for all $z \in \mathscr{Z}$.

(e) $(X_s|Z, R=1) \overset{d}{=} (X_a|Z, R=0)$.

Assumption 2.1(a) is a support condition on the commonly observed variables. It ensures that we will be able to find, for all the observations in the study sample, comparable units in the auxiliary sample, Assumption 2.1(b) imposes a pseudo randomization scheme on R, and therefore, allows us to view the merged data as a random sample from the pseudo-merged population. Let $\ell(\cdot)$ denote the conditional likelihood ratio of Z across two population, i.e. $\ell(z) \equiv f_{Z|R}(z|1)/f_{Z|R}(z|0)$. Assumption 2.1(c) expresses this likelihood ratio as a function of $r(\cdot)$, which plays the role of the "propensity score" function of R given Z. In our context, this is the probability that one observation belongs to the study population conditional on the value that instrumental variables take. The first part of Assumption 2.1(d) indicates that n_s grows at the same order of magnitude as n_a. The second part of Assumption 2.1(d) ensures that the pseudo-true merged population is not a degenerate one conditional on all possible values of Z. By Assumption 2.1(b)–(d) and Bayes' Law, we have $r(z) = \mathbb{P}(R = 1|Z = z)$, and thus, $r(\cdot)$ can be viewed as the propensity score function.

Assumption 2.1(e) is a rank similarity condition. It requires the conditional distribution of X_s given Z in the principal population coincide with that of X_a in the auxiliary population. In view of the structural relation in (2.2.2), the assumption is satisfied if $h_s = h_a$ and $(\eta_s|Z, R = 1) \overset{d}{=} (\eta_a|Z, R = 0)$. Assumption 2.1(e) is the only cross-population restriction we impose on our data structure, which means the conditional distribution of Y given (X, Z), and therefore, the conditional distribution of Y given Z and the marginal distribution of Z are all allowed to differ across $\mathscr{P}_s$ and $\mathscr{P}_a$. This assumption is weaker than Assumption 1(ii) of Graham et al. (2016), as we do not impose a rank similarity condition on the outcome, which would imply $F_{Y_s|ZR=1}(\cdot|\cdot) = F_{Y_a|ZR=0}(\cdot|\cdot)$.

2.3 Identification

In this section, we first introduce the definition of UQE. Then, we develop a set of identification results, for the cases when X is continuously distributed, and when it is discrete, respectively.

2.3.1 Parameter of Interest

Our definition of the unconditional policy effect depends on the notion of a counterfactual experiment, which is formally defined as follows:

Definition 2.1 (Counterfactual Experiment) Let $\phi \equiv (\widetilde{\mathscr{U}_s}, \widetilde{G}_s, \widetilde{Z}, \widetilde{R}, \widetilde{\varepsilon}_s, \widetilde{g}_s) : \Omega \to K([0,1]) \times \mathscr{D}(\mathscr{X}) \times \mathscr{Z} \times \{0,1\} \times \mathscr{E} \times l_2(\mathscr{X}, \mathscr{Z}_1, \mathscr{E})$, where $K([0,1])$ is the collection of all non-empty closed subsets of the unit interval, and $\mathscr{D}(\mathscr{X})$ denotes the space of distribution functions on $\mathscr{X}$. We say Φ is the set of counterfactual experiments, if for all $\phi \in \Phi$, we have (i) $\widetilde{G}_s^{-1}(U_s) = \widetilde{G}_s^{-1}(U_s')$ almost surely for all $U_s, U_s' \in \widetilde{\mathscr{U}_s}$; (ii) $(\widetilde{\varepsilon}_s, \widetilde{Z}, \widetilde{R}) \overset{d}{=} (\varepsilon_s, Z, R)$; (iii) $\widetilde{g}_s = g_s$, (iv) for all $U_s \in \mathscr{U}_s$ and $\widetilde{U}_s \in \widetilde{\mathscr{U}_s}$, there exists $\widetilde{U}_s' \in \widetilde{\mathscr{U}_s}$ and $U_s' \in \mathscr{U}_s$, respectively, such that $(\widetilde{U}_s'|\widetilde{Z}_1, \widetilde{R} = 1) \overset{d}{=} (U_s|Z_1, R = 1)$ and $(\widetilde{U}_s|\widetilde{Z}_1, \widetilde{R} = 1) \overset{d}{=} (U_s'|Z_1, R = 1)$, where $\mathscr{U}_s = \{\breve{U} \in \mathscr{U}[0,1] : (F_{X_s|R}^{-1}(\breve{U}_s|1)|Z_1, R = 1) \overset{d}{=} (X_s|Z_1, R = 1)\}$.

The definition of counterfactual experiments does not specify the counterfactural target covariate $\widetilde{X}_s$ directly. It is implicitly defined through the first two elements of ϕ. The first element, $\widetilde{\mathscr{U}_s}$, is a set of rank variables associated with the counterfactual target covariate, $\widetilde{X}_s$. When $\widetilde{X}_s$ is absolutely continuous, $\widetilde{\mathscr{U}_s}$ becomes a singleton set, but the set is generally not degenerate when the distribution of $\widetilde{X}_s$ contains a mass point. The second component, $\widetilde{G}_s$, is the counterfactual distribution of $\widetilde{X}_s$ conditional on the study population. When the target covariate is continuously distributed, $\widetilde{G}_s$ is continuous and strictly increasing, and therefore, $\widetilde{X}_s$ is uniquely determined by $\widetilde{X}_s = \widetilde{G}_s^{-1}(\widetilde{U}_s)$, where $\widetilde{U}_s$ is the only element in $\widetilde{\mathscr{U}_s}$. However, when the target covariate contains mass points, there is a set of counterfactual rank variables that correspond to the same target covariate in the study population. This equivalent class is defined by Condition (i).

Following Rothe (2012), we restrict our attention to counterfactual changes where only the marginal distribution of X_s is changed, while the marginal distribution of Z and the dependence structure between X_s and Z remain unaffected. This notion of a *ceteris paribus* change is formally characterized by Conditions (ii)–(iv). Condition (ii) implies that the joint distribution of the observed variables (Z,R) and the latent variable ε_s remain unchanged across counterfactual experiments. Under Condition (iii), the structural production function, g is also not affected by the counterfactual change. Condition (iv) imposes a rank similarity condition. It says the conditional rank of the counterfactural target covariate follows the same distribution as the status quo. Due to the possibility of multiplicity of rank variables, the condition is also framed in terms of a set equivalence condition. When we restrict attention to absolutely continuous target covariates, both $\mathscr{U}_s$ and $\widetilde{\mathscr{U}_s}$ are singleton sets. Hence, this condition reduces to $(\widetilde{X}_s|\widetilde{Z}_1,\widetilde{R}=1) \overset{d}{=} (X_s|Z_1,R=1)$.

Each counterfactual experiment ϕ represents a modification of the underlying economic system. It completely determines the counterfactual outcome in the study population. Yet we remain largely agnostic as to the counterfactual change in the auxiliary population. The definition also leaves the mechanism causing the change in the marginal distribution of the target covariate unspecified.

Remark 2.1 Our definition of counterfactual experiments relaxes the rank invariance conditions imposed by Rothe (2012). Instead, counterfactual changes in our context only need to satisfy a rank similarity or copula invariance condition.

With the counterfactual experiments defined, we now construct the counterfactual covariate vector by $\widetilde{W}_G = (\widetilde{G}_s^{-1}(\widetilde{U}_s),\widetilde{Z}_1')'$. The counterfactual outcome of the study population is then defined as $\widetilde{Y}_s = \widetilde{g}_s(\widetilde{W}_G,\widetilde{\varepsilon}_s)$, which follows a marginal distribution, $F_{\widetilde{Y}_s}$, and a conditional distribution restricted to the principal population, $F_{\widetilde{Y}_s|R=1}$. Note that the unconditional distribution is not well-defined, due to the lack of information on counterfactual changes in the auxiliary population. Therefore, we focus exclusively on the counterfactual distribution conditional on the study population in what follows. When X is discrete, a single counterfactual experiment is mapped to a set of counterfactual outcomes, and we denote the corresponding set of counterfactual distributions by $\mathscr{F}_{\widetilde{Y}_s}$.

In our context, the sequence of counterfactual distributions is defined in terms of the "marginal" distribution of the potential covariate X_s in the study population, rather than the true unconditional distribution of the observed X. Although X_s is missing from the main dataset, and therefore, its marginal distribution cannot be directly identified from the study population, we show in Theorem 2.1 that it can be recovered from the auxiliary data under the rank similarity assumption we impose in Assumption 2.1.

The policy parameter we seek to identify in this paper is the pathwise derivative of counterfactual distributional effect conditional on the study population. It is adapted from the definition of the *marginal partial distributional policy effect* (MPPE) by Rothe (2012).

Definition 2.2 (Marginal Partial Distributional Policy Effect) Let $\Phi^* \equiv \{\phi_t\}_{t\geq 0} \subset \Phi$ denote a sequence of counterfactual experiments, such that $\widetilde{G}_{s,t} \to F_{X_s|R=1}$, as $t \downarrow 0$. The MPPE for a given functional $\nu : \mathscr{D}(\mathscr{Y}) \to \mathbb{R}$ and a sequence of $\widetilde{F}_{s,t} \in \mathscr{F}_{\widetilde{Y}_{s,t}}$ is defined by,

$$MPPE(\nu, \{\widetilde{Y}_{s,t}\}_{t\geq 0}) \equiv \left.\frac{\partial \nu(F_{\widetilde{Y}_{s,t}|\widetilde{R}=1})}{\partial t}\right|_{t=0} = \lim_{t\downarrow 0} \frac{\nu(F_{\widetilde{Y}_{s,t}|\widetilde{R}=1}) - \nu(F_{Y_s|R=1})}{t}.$$

We consider two specific types of counterfactual distributional changes: MDS and MQS. The defintion of the former is due to Firpo et al. (2009b). It denotes a small perturbation in the distribution of X_s, in the direction of G. MQS, on the other hand, considers a minuscule change in the quantiles of X_s. This type of policy change includes the MLS, $G^{-1}_{t,ls}(u) \equiv F^{-1}_{X_s|R}(u|1) + t$, as a special case.

Definition 2.3 (Counterfactual Policy Distributions)

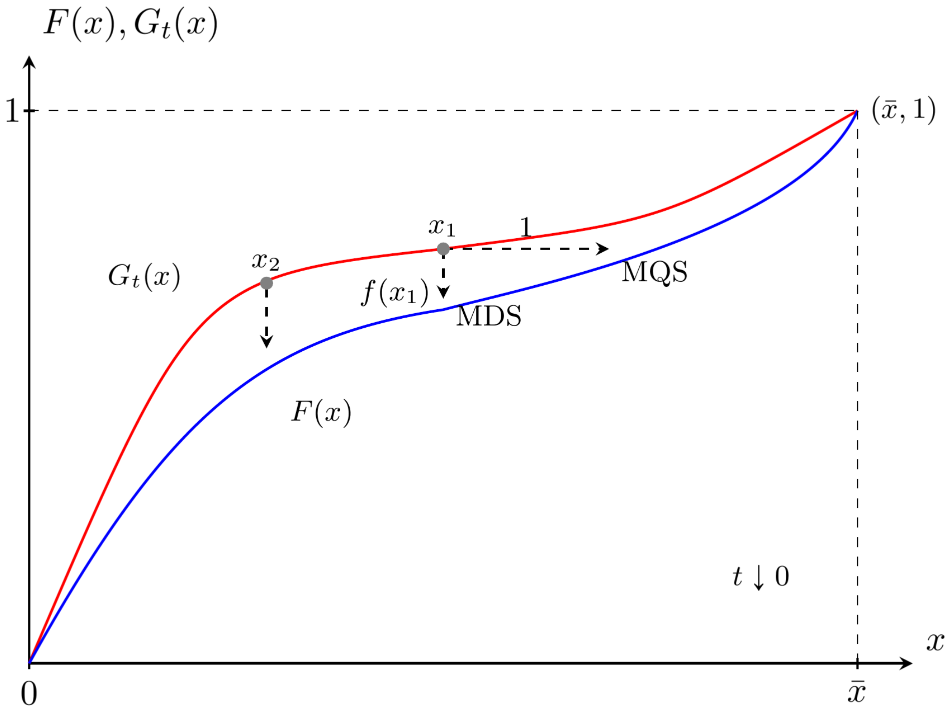
Marginal Distributional Shift (MDS): $G_{t,p}(x) = F_{X_s|R}(x|1) + t(G(x) - F_{X_s|R}(x|1))$.

Marginal Quantile Shift (MQS): $G^{-1}_{t,q}(u) = F^{-1}_{X_s|R}(u|1) + t(G^{-1}(u) - F^{-1}_{X_s|R}(u|1))$.

Remark 2.2 Figure 2.1 illustrates how the rates of change between the two types of counterfactuals are related. Under the condition that $F_{X_s|R=1}$ is compactly supported with strictly positive density on $\mathscr{X}$, MQS in a user-specified direction, $q(x)$, can be approximated in the limit by MDS with $G(x) = F_{X_s|R}(x|1) - f_{X_s|R}(x|1)q(x)$.

Turning to the case of quantiles, the quantile operator for a particular τ is defined by, $\nu_\tau(F_{Y_s|R=1}) = F^{-1}_{Y_s|R=1}(\tau)$. With the understanding that MPPE associated with a counterfactual experiment is generally a set when the X is discretely valued, we suppress the index with respect to $\{\widetilde{Y}_{s,t}\}_{t\geq 0}$ for notational convenience, and denote the MPPE with MDS, $MPPE(\nu_\tau, \{\widetilde{Y}_{s,t}\}_{t\geq 0})$, and MPPE with MQS, $MPPE(\nu_\tau, \{\widetilde{Y}_{s,t}\}_{t\geq 0})$, by $UQE_p(\tau, G)$ and $UQE_q(\tau, G)$, respectively. Here, and in what follows, the qualifier "unconditional" in UQE should be understood as conditional on (or relative to) the study population.

Figure 2.1: Marginal distributional shift and marginal quantile shift



Notes: The blue curve depicts the data distribution of X. The red curve depicts a counterfactual distribution in the sequence $\{G_t\}_{t\downarrow 0}$, which can be induced by two equivalent policy changes.

2.3.2 First Step Identification

If there is no missing variable, the joint distribution of (Y, X, Z) is directly identifiable from a random sample. Under data combination, however, only the "marginal" conditional distributions: $F_{Y_s|ZR=1}$ and $F_{X_a|ZR=0}$, can still be separately identified from the two samples, respectively. The conditional distribution, $F_{Y_s|X_sZ_1R=1}$, is generally not identifiable without further cross-population assumptions.

Instead of seeking identification of the entire conditional distribution, $F_{Y_s|X_sZ_1R=1}(\cdot|\cdot,\cdot,1)$, we demonstrate in Sections 2.3.3 and 2.3.4 that UQE, and MPPE in general, can be identified using information on a finite set of points of Y_s. For any τ that belongs to this set, we define q_τ as the $\tau-$th quantile of Y_s. Choice of τ depends on research interest. For instance, it can include only the median, the quartiles of Y_s, etc. This flexibility allows us to obtain identification under much milder restrictions on the pseudo-merged population.

Identification is achieved through the excluded instrument variables, Z_2. To ease notational burden, we define $\Lambda(x, z_1) \equiv F_{Y_s|X_sZ_1R}(q_\tau|x, z_1, 1)$ for a given τ and for all $x \in \mathscr{X}$ and $z_1 \in \mathscr{Z}_1$.

Assumption 2.2 $\varepsilon_s \perp\!\!\!\perp Z_2 | X_s, Z_1, R = 1$.

Assumption 2.2 implies that Z_2 can be excluded from the outcome equation, and therefore, can be used as a source of exogenous variation to proxy for the missing covariate in the study population. Note that Λ is generally not identified without an exogenous instrument Z_2. We illustrate this point with a linear normal model in Example 2.1.

Under Assumptions 2.1 and 2.2, the following moment-matching equation holds,

$$\mathbb{E}[1(Y \leq q_\tau)|Z, R=1] - \mathbb{E}[\Lambda(W)|Z, R=0] = 0, \text{ a.s.} \tag{2.3.1}$$

Equivalently, Λ can be identified based on a likelihood-ratio-weighting equation,

$$\mathbb{E}\left[R1(Y \leq q_\tau) - (1-R)\Lambda(W)\frac{r(Z)}{1-r(Z)} \middle| Z \right] = 0, \text{ a.s.} \tag{2.3.2}$$

The next assumption is about the global identification of Λ.

Assumption 2.3 Λ is the unique solution to (2.3.1) or (2.3.2) almost surely.

Assumption 2.3 is a high level condition. It is implied by a bounded completeness condition on the auxiliary population.[2] Note that Λ is globally identified as long as $\mathbb{E}[\Lambda(W) - \widetilde{\Lambda}(W)|Z, R=0] = 0$ implies $\Lambda = \widetilde{\Lambda}$, which follows immediately if Λ is measurable with respect to W, and W is bounded complete for Z, relative to the auxiliary population.[3] Although Canay et al. (2013) show that the completeness condition is untestable against general alternatives, we use two examples to show that the assumption is reasonable in some special cases.

The completeness condition implicitly imposes some constraints on the support of the excluded instrument. Whenever X is continuously distributed, Z_2 is generally required to be continuous. In Example 2.1, we show that when both X and Z_2 are continuous, Assumption 2.3 holds when the structural errors follow a joint normal distribution, which is a commonly-adopted assumption in empirical practices. However, our method does not apply if the instrument has finite support or otherwise violates the bounded completeness condition, the latter of which is likely to occur if the strength of the instrument is weak.

On the other hand, when X is discretely valued, we show via Example 2.2 that Assumption 2.3 can be satisfied with a discrete instrument. The key requirement is a rank condition on conditional probability matrices of X_a given Z. When the set of Z_1 is empty, we can uses Cragg and Donald (1996) or Robin and Smith (2000) to test the rank condition.

[2] For two random element U and V, we say U is bounded complete for V, relative to a subpopulation $S = s$, if for all bounded measurable functions $\delta(\cdot)$, $\mathbb{E}[\delta(U)|V, S = s] = 0$ implies $\delta(U) \equiv 0$ almost surely.

[3] Bounded completeness is weaker than the commonly adopted completeness condition appearing in Newey and Powell (2003) and Fan et al. (2014). We refer readers to Hoeffding et al. (1977), Blundell et al. (2007), and Lehmann (1986) for detailed discussions.

In Section 2.4, we base our estimation and inference on parametric identification of Λ. In this case, we assume that $F_{Y_s|X_sZ_1R}(q_\tau|x,z_1,1) = \Lambda(x,z_1;\beta_0)$, for some $\beta_0 \in \theta_\beta \subset \mathbb{R}^{d_\beta}$. In Lemma 2.3, we provide a set of sufficient conditions which allow us to establish a global parametric identification condition analogous to Assumption 2.3.

Lemma 2.1 $F_{Y_s|X_sZ_1R}(q_\tau|\cdot,\cdot,1)$ is point identified under Assumptions 2.1–2.3.

Lemma 2.1 establishes the nonparametric identification of Λ. The proof for the parametric case follows along exactly the same line so we omit it here. In the next example, we verify the identification assumptions in a conditional normal model.

Example 2.1 (Conditional Normal Model) Let the structural equations of the study population be given by

$$
\begin{aligned}
Y_s &= g_s(X_s,Z_1) + \varepsilon_s, \\
X_s &= h_s(Z) + \eta_s,
\end{aligned}
$$

where ε_s and η_s are jointly normally distributed. Specifically, for positive-valued functions $\psi_y(\cdot)$ and $\psi_x(\cdot)$, we have

$$
(\varepsilon_s,\eta_s)|Z,R=1 \sim N\left(0, \begin{pmatrix} \psi_y(Z_1) & 0 \\ 0 & \psi_x(Z_1) \end{pmatrix}\right).
$$

Then, $\Lambda(w) = \Phi(\psi_y(z_1)^{-1/2}(q_\tau - g_s(w)))$, where $\Phi(\cdot)$ denotes the CDF of standard normal distribution. Suppose the reduced-form of X given Z in the auxiliary population is $X_a = h_a(Z) + \eta_a$, Assumption 2.1(e) is satisfied if $h_s = h_a = h$, and $(\eta_s|Z,R=1) \overset{d}{=} (\eta_a|Z,R=0)$. Assumption 2.2 holds if $Z_2 \perp\!\!\!\perp \varepsilon_s|X_s,Z_1,R=1$. Assume, in addition that, conditional on z_1, $\text{Supp}(F_{Z_2})$ contains an open set and that $h(z_1,\cdot)$ maps open sets of z_2 into open sets. Assumption 2.3 then follows by Theorem 2.2 in Newey and Powell (2003).

Turning to the linear case, let $g_s(w) = \gamma_{s_1}x + \gamma_{s_2}'z_1$, $h(z) = \delta_1'z_1 + \delta_2'z_2$, $\psi_y = \psi_x = 1$, $\eta_s \perp\!\!\!\perp (\varepsilon_s, Z_2)$, and therefore, $\mathbb{E}[1(Y \leq q_\tau)|Z,R=1] = \Phi((q_\tau - (\gamma_{s_1}\delta_1' + \gamma_{s_2})'Z_1 - \gamma_{s_1}\delta_2'Z_2)/(1+\gamma_{s_1}^2)^{1/2})$. As a consequence, $(\gamma_{s_1},\gamma_{s_2}')'$ are uniquely determined by (2.3.1) or (2.3.2), if and only if $\delta_2 \neq 0$.

Example 2.2 (Discrete Covariates) Suppose X_s, X_a, and Z_2 are all discretely valued. Assume that $\text{Supp}(F_{X_j|Z_1=z_1}) = \{x^1,...,x^l\}$ and $\text{Supp}(F_{Z_2|Z_1=z_1}) = \{z^1,...,z^k\}$, for $j = s,a$ and for all $z_1 \in \text{Supp}(F_{Z_1})$. Let $P^j_{u,t}(z_1,r) = \mathbb{P}(X_j = x^u|Z_2 = z^t, Z_1 = z_1, R = r)$ for $j = s,a$ and $r = 0,1$. Assumption 2.1(e) holds if for all $u \in \{1,2,...,l\}$ and $t \in \{1,2,...,k\}$, $P^s_{u,t}(Z_1,1) = P^a_{u,t}(Z_1,0)$ with probability 1. Moreover, let $P^j(\cdot,\cdot)$ denote the matrix of probabilities where the (u,t)-th entry is equal to $P^j_{u,t}(\cdot,\cdot)$. Then by Theorem 2.4 in Newey and Powell (2003), bounded completeness, and hence, Assumption 2.3, holds if $rank(P^a(Z_1,0)) = l$ with probability 1.

2.3.3 Identification with Continuously Distributed Covariates

In this section, we establish the identification of UQE_q and UQE_p when the distribution of X is absolutely continuous. Before stating the main result, we need some additional identifying assumptions.

Assumption 2.4

(a) (i) $\varepsilon_s \perp\!\!\!\perp U_s | Z_1, R = 1$; (ii) there exists a t_0 sufficiently close to 0, such that for all $t \leq t_0$ and $\phi_t \in \Phi^*$, $\widetilde{\varepsilon}_{s,t} \perp\!\!\!\perp \widetilde{U}_{s,t} | \widetilde{Z}_{1,t}, \widetilde{R}_t = 1$.

(b) $\text{Supp}(G) \subset \text{Supp}(F_{X|ZR}(\cdot|Z,1))$ almost surely.

Assumption 2.5 $F_{Y|R=1}$ is continuously differentiable in an open neighborhood of q_τ with strictly positive density function $f_{Y|R=1}$.

Assumption 2.4(a)(i) imposes that conditional on Z_1, structural error ε_s is independent of the rank variable U_s in the study population. This is much weaker than the commonly assumed strict independence condition that X_s is independent of ε_s unconditionally. Conditional exogeneity has also been imposed by Firpo et al. (2009b), Rothe (2012), and Chernozhukov et al. (2013a), among others. Assumption 2.4(a)(ii) requires the conditional independence condition of part (a) to hold when counterfactual experiments get sufficiently close to the status quo. Under the rank invariance condition imposed by Rothe (2012), it is automatically implied by Assumption 2.4(a)(i). Assumption 2.4(b) ensures that the conditional distribution of Y_s given W is identified over the support of W. Assumption 2.5 imposes a smoothness condition on the distribution of target outcome, which implies that $F^{-1}_{Y|R=1}$ is Hadamard differentiable at $F_{Y|R=1}$, tangentially to the set of functions that are continuous at q_τ.

The main theoretical result of this section is given as follows:

Theorem 2.1 Suppose that Assumptions 2.1–2.5 hold, and that the distribution of X is absolutely continuous with respect to the Lebesgue measure, both $UQE_p(\tau, G)$ and $UQE_q(\tau, G)$ are identified.

(a) For UQE_q, we have

$$UQE_q(\tau, G) = -\frac{1}{f_{Y_s|R}(q_\tau|1)(1-Q_0)} \mathbb{E}\left[(1-R)\ell(Z)\Lambda_x(X, Z_1)g_q(X)\right],$$

where $g_q(x) \equiv G^{-1}(F_{X_s|R}(x|1)) - x$, $\Lambda_x(x, z_1) \equiv \partial\Lambda(\tilde{x}, z_1)/\partial\tilde{x}|_{\tilde{x}=x}$, and $F_{X_s|R}(x|1) = \frac{1}{1-Q_0}\mathbb{E}\left[(1-R)\ell(Z)1(X \leq x)\right]$.

(b) Suppose in addition that $\mathscr{X}$ is compact, and $F_{X_s|R=1}$ is continuously differentiable on $\mathscr{X}$ with strictly positive density function $f_{X_s|R=1}$. Then we have,

$$UQE_p(\tau, G) = -\frac{1}{f_{Y_s|R}(q_\tau|1)(1-Q_0)} \mathbb{E}\left[(1-R)\ell(Z)\Lambda_x(X, Z_1)g_p(X)\right],$$

where $g_p(x) \equiv -\frac{G(x)-F_{X_s|R}(x|1)}{f_{X_s|R}(x|1)}$, and $f_{X_s|R}(x|1) = \frac{1}{1-Q_0}\partial\mathbb{E}\left[(1-R)\ell(Z)1(X \leq x)\right]/\partial x$.

Remark 2.3 The compactness condition on $\mathscr{X}$ is assumed to ensure the existence of pathwise derivative of the inverse map. It can be relaxed by imposing a boundary condition on Λ_x. Specifically, we may assume that Λ_x vanishes when $x \notin [F_{X_s|R=1}(q_1)+\varepsilon, F_{X_s|R=1}(q_2)-\varepsilon]$, for $0 < q_1 < q_2 < 1$ and some $\varepsilon > 0$.

2.3.4 Identification with Discrete Covariate

Let the support of X be $\{x^1,\dots,x^l\}$. When X is discrete, MQS is not well-defined and we consider MDS only, with counterfactual experiments defined through a fixed discrete distribution, G. As indicated by Example 2.2, results in this section hold when Z_2 are both continuously and discretely valued.

Assumption 2.6

(a) (i) $\varepsilon_s \perp\!\!\!\perp U_s|Z_1, R=1$, for all $U_s \in \mathscr{U}_s$; (ii) there exists a t_0 sufficiently close to 0, such that for all $t \leq t_0$ and $\phi_t \in \Phi^*$, $\varepsilon_{s,t} \perp\!\!\!\perp \widetilde{U}_{s,t}|\widetilde{Z}_{1,t},\widetilde{R}_t = 1$, for all $\widetilde{U}_{s,t} \in \widetilde{\mathscr{U}_{s,t}}$.

(b) $\mathrm{Supp}(G) \subset \mathrm{Supp}(F_{X_s|R=1})$.

(c) For all $U_s \in \mathscr{U}_s$, $F_{U_s|Z_1R}(u_s|z_1,1)$ is continuously differentiable in u_s, for all $z_1 \in \mathscr{Z}_1$.

Assumption 2.6(a) is the counterpart of Assumption 2.4(a) for discrete covariates. Since the rank variables are no longer uniquely pinned down by strictly increasing quantile functions, we strengthen Assumption 2.4(a) so that conditional independence holds for all the rank variables in the equivalent class. With this identifying assumption in hand, we are ready to present the following identification result. For $j = 1,\dots,l$, let the period bound generating function be defined by

$$h_{q_\tau}(x^j, x^{j-1}, z_1) \equiv -\frac{(\Lambda(x^{j-1},z_1)-\Lambda(x^j,z_1))\cdot(G(x^{j-1})-F_{X_s|R}(x^{j-1}|1))}{f_{Y_s|R}(q_\tau|1)}.$$

Theorem 2.2 Suppose that Assumptions 2.1–2.3, 2.5, and 2.6 hold, $UQE_p(\tau,G)$ is partially identified, with

$$UQE_p(\tau,G) \in \Bigg[\sum_{j\in\mathscr{J}_+} h_{q_\tau}(x^j,x^{j-1},z^{\dagger}_{1,j}) + \sum_{j\in\mathscr{J}_-} h_{q_\tau}(x^j,x^{j-1},z^{*}_{1,j}),$$
$$\sum_{j\in\mathscr{J}_+} h_{q_\tau}(x^j,x^{j-1},z^{*}_{1,j}) + \sum_{j\in\mathscr{J}_-} h_{q_\tau}(x^j,x^{j-1},z^{\dagger}_{1,j})\Bigg],$$

where $\mathscr{J}_+ \equiv \{j \in \{1,\dots,l\} : G(x^{j-1}) \leq F_{X_s|R}(x^{j-1}|1)\}$ ($\mathscr{J}_-$ is analogously defined), $z^*_{1,j} \equiv \arg\sup_{z_1\in\mathscr{Z}_1}(\Lambda(x^{j-1},z_1)-\Lambda(x^j,z_1))$, $z^{\dagger}_{1,j} \equiv \arg\inf_{z_1\in\mathscr{Z}_1}(\Lambda(x^{j-1},z_1)-\Lambda(x^j,z_1))$, and $F_{X_s|R}(x^j|1) = \mathbb{E}[\frac{1-R}{1-Q_0}\ell(Z)1(X \leq x^j)]$, for $j \in \{1,\dots,l\}$.

Theorem 2.2 indicates that UQE_p is generally partially identified with bounds generated by h_{q_τ}. In the special case when $\Lambda(x,z_1)$ is constant in z_1, the identified set of UQE_p reduces to a singleton.

If X is binary and $G_{t,p}(x) = 1\{0 \leq x < 1\}(F_{X_s|R}(0|1) - t) + 1\{x \geq 1\}$, h_{q_τ} reduces to $-(\Lambda(1,z_1) - \Lambda(0,z_1)) / f_{Y|R}(q_\tau|1)$. In such circumstance, Theorem 2.2 corresponds to the two-sample generalization of Theorem 5 in Rothe (2012), when ν in that paper takes on the quantile functional.

2.4 Estimation and Inference

In this section, we discuss estimation and inference for our two-sample UQE. First, we describe an estimation procedure for UQE_q and UQE_p as identified in Theorem 2.1. We then show that our estimator is consistent and asymptotically normal in Theorem 2.3.[4]

2.4.1 Estimation Procedure

Following the discussion in Section 2.3, we first propose an estimator of the conditional probability, Λ. Here, we restrict our attention to the parametric setting where Λ is indexed by a vector of parameter, β. We use a moment-matching method based on (2.3.2) to estimate β. The estimation of β consists of four-steps. In the first step, we estimate q_τ by solving

$$\widehat{q}_\tau \equiv \arg\min_{q \in \mathscr{Y}} \mathbb{E}_n[R(\tau - 1(Y \leq q)) \cdot (Y - q)]. \tag{2.4.1}$$

The next three steps follow closely the Auxilliary-to-Study Tilting (AST) method proposed by Graham et al. (2016). Using the AST estimator, β and the propensity score can be jointly estimated from moment restrictions in (2.3.2). To implement the estimator, we first estimate the propensity score, $r(z)$. Towards this end, we assume that the propensity score takes a parametric form, i.e. $r(z) = L(k(z)'\gamma)$, where $L(\cdot)$ is any link function that satisfies Assumption 2.7(e). Using $L(\cdot)$, $\widehat{\gamma}$ can be obtained by solving the following problem,

$$\widehat{\gamma} \equiv \arg\max_{\gamma \in \Theta_\gamma} \mathbb{E}_n \left[R \log(L(k(Z)'\gamma)) + (1 - R)\log(1 - L(k(Z)'\gamma)) \right]. \tag{2.4.2}$$

The AST estimator augments the conditional maximum likelihood estimator $\widehat{\gamma}$ with tilting parameters. The resulting estimator of β is more efficient than the one based on $\widehat{\gamma}$ alone. Let $t(z)$ be a vector of known functions of z with a constant term as the first element. Denote the tilting parameters associated with the auxiliary data and the study

[4] Here we focus on the scenario where the distribution of X is absolutely continuous. When X is discrete, the problem features partially identified parameters defined by the intersection bounds. Chernozhukov et al. (2013b) provide an extensive treatment of this topic. We omit discussion here and refer readers to Appendix D in Rothe (2012) for a detailed discussion on how to apply their method.

sample, by λ_a and λ_s, respectively. They are estimated by solving,

$$\mathbb{E}_n\left[\left(\frac{1-R}{1-L(k(Z)'\widehat{\gamma}+t(Z)'\widehat{\lambda}_a)}-1\right)L(k(Z)'\widehat{\gamma})t(Z)\right]=0, \tag{2.4.3}$$

$$\mathbb{E}_n\left[\left(\frac{R}{L(k(Z)'\widehat{\gamma}+t(Z)'\widehat{\lambda}_s)}-1\right)L(k(Z)'\widehat{\gamma})t(Z)\right]=0. \tag{2.4.4}$$

Using $\widehat{\lambda}_s$ and $\widehat{\lambda}_a$, we compute study and auxiliary sample tilts, which are defined as follows:

$$\widehat{\pi}_i^s \equiv \frac{L(k(Z_i)'\widehat{\gamma})}{L(k(Z_i)'\widehat{\gamma}+t(Z_i)'\widehat{\lambda}_s)}, \qquad \widehat{\pi}_i^a \equiv \frac{L(k(Z_i)'\widehat{\gamma})}{1-L(k(Z_i)'\widehat{\gamma}+t(Z_i)'\widehat{\lambda}_a)}. \tag{2.4.5}$$

Also let $e(z)$ be a d_β-dimensional vector of known functions of z, and $g(a;\widehat{q}_\tau,\widehat{\gamma},\widehat{\lambda}_s,\widehat{\lambda}_a,\beta)\equiv(\widehat{\pi}^s r1(y\leq\widehat{q}_\tau)-\widehat{\pi}^a(1-r)\Lambda(w;\beta))e(z)$. Now, in the last step, β can be estimated by

$$\widehat{\beta}\equiv\arg\inf_{\beta\in\Theta_\beta}\widehat{\mathscr{L}_n}(\beta), \tag{2.4.6}$$

where $\widehat{\mathscr{L}_n}(\beta)\equiv\left\|\mathbb{E}_n[g(A;\widehat{q}_\tau,\widehat{\gamma},\widehat{\lambda}_s,\widehat{\lambda}_a,\beta)]\right\|^2_{\Omega_n}$ and $\|x\|^2_{\Omega_n}\equiv x'\Omega_n x$, for a sequence of positive definite weighting matrices Ω_n.

Using these quantities, we can obtain $\Lambda_x(w;\widehat{\beta})\equiv\partial\Lambda(\tilde{x},z_1;\widehat{\beta})/\partial\tilde{x}|_{\tilde{x}=x}$, and $\widehat{\ell}(z)\equiv\frac{n_a}{n_s}\cdot\frac{L(k(z)'\widehat{\gamma}+t(z)'\widehat{\lambda}_s)}{1-L(k(z)'\widehat{\gamma}+t(z)'\widehat{\lambda}_a)}$. Throughout this section, we assume that the counterfactual distribution G is known. In practice, if G is not known, it may be estimated from an independent sample; see e.g. Rothe (2010). Using the above estimates and $\widehat{F}_{X_s|R=1}(\cdot)\equiv\mathbb{E}_{n_a}[\widehat{\ell}(z)1(X\leq\cdot)]$, where $\mathbb{E}_{n_a}[X]$ denotes $n_a^{-1}\sum_{i=n_s+1}^{n}X_i$, $\widehat{g}_q$ can be obtained as the plug-in estimator. For g_p, we need an estimator for $f_{X|R=1}(\cdot)$. Our identification relies on a compact support condition, and it is well known that the Prazen-Rosenblatt density estimator is not valid near the boundary of support. To overcome this challenge, we introduce trimming.[5] For a kernel K_x with compact support, and some bandwidth b_x, we let

$$\widehat{f}_{X|R}(x|1)\equiv\mathbb{E}_{n_a}\left[\widehat{\ell}(Z)I_{b_x}K_{b_x}(X-x)\right],$$

where $K_{b_x}(\cdot)\equiv b_x^{-1}K_x(\cdot/b_x)$. I_{b_x} is a trimming indicator, which equals one for $x\in\{[\underline{x}+\rho_x b_x/2,\bar{x}-\rho_x b_x/2]\}$, where $\underline{x}$, $\bar{x}$, and ρ_x are the lower and upper bound, of $\mathscr{X}$, and the diameter of $\text{Supp}(K_x)$, respectively. The density, $f_{Y|R}$, can also be estimated using a kernel density estimator. Specifically, for any kernel function $K_y(\cdot)$ that satisfies Assumption

[5] Trimming is widely adopted in the literature; see e.g. Härdle and Stoker (1989), Powell et al. (1989) among others. This specific trimming function is inspired by Guerre et al. (2000) and Li et al. (2002). As an alternative, we can use a local polynomial density estimator that adjusts for the boundary bias adaptively; see Cattaneo et al. (2020) for details.

2.8(b), let $\widehat{f}_{Y|R}(y|1) \equiv \mathbb{E}_{n_s}[K_{b_y}(Y_i - y)]$, where $\mathbb{E}_{n_s}[X] \equiv n_s^{-1}\sum_{i=1}^{n_s} X_i$ and $K_{b_y}(y) \equiv b_y^{-1}K_y(y/b_y)$.

Now, plugging in the estimators of nuisance quantities, $UQE_j(\tau, G)$ can thus be estimated by,

$$\widehat{UQE}_j(\tau, G) \equiv -\frac{1}{\widehat{f}_{YR}(\widehat{q}_\tau|1)}\mathbb{E}_{n_a}[\widehat{\ell}(Z)\Lambda_x(W;\widehat{\beta})\widehat{g}_j(X)],\ j = p, q. \tag{2.4.7}$$

We summarize the estimation procedure in the following algorithm.

Algorithm 2.4.1 (Plug-in Estimator for $\widehat{UQE}$)

1. Compute the empirical quantile estimator $\widehat{q}_\tau$ by solving (2.4.1).

2. Compute the conditional maximum likelihood estimator $\widehat{\gamma}$ by solving (2.4.2).

3. Solve (2.4.3) and (2.4.4) to get $\widehat{\lambda}_j$, and use them to compute $\widehat{\pi}_j$, for $j = s, a$, following (2.4.5).

4. Use the above quantities to compute $\widehat{\beta}$, by solving (2.4.6).

5. Compute $\Lambda_x(\cdot;\widehat{\beta}), \widehat{\ell}(\cdot), \widehat{F}_{X_s|R=1}(\cdot)$, $\widehat{f}_{X_s|R=1}(\cdot)$. Using these quantities to compute $\widehat{g}_j$, for $j = p, q$.

6. For $j = p, q$, compute the plug-in estimator $\widehat{UQE}_j$ following (2.4.7).

2.4.2 Large Sample Results

In this section, we present inference results for the estimators introduced in the previous section. We first establish large sample properties of $\widehat{\beta}$, for which purpose, some additional regularity conditions are in order.

Assumption 2.7

(a) (i)$\{(R_i, R_iY_i, (1-R_i)X_i, Z_i)\}_{i=1}^n$ are *i.i.d.*; (ii) let $\theta \equiv (\gamma, \lambda_s, \lambda_a, \beta) \in \Theta \equiv \Theta_\beta \times \Theta_\lambda^2 \times \Theta_\beta$, then Θ is compact, and θ_0 lies in the interior of Θ.

(b) $F_{Y|ZR=1}(y|z)$ is absolutely continuous and differentiable in $y \in \mathscr{Y}_0$ for all $z \in \mathscr{Z}$, where $\mathscr{Y}_0$ is a compact subset of $\mathscr{Y}$, and

$$\sup_{(y,z)\in\mathscr{Y}_0\mathscr{Z}} |f_{Y|ZR}(y|z,1)| \leq c_1 < \infty.$$

(c) (i) $\Lambda(w;\beta)$ is twice continuously differentiable in β with uniformly bounded derivatives, for all $w \in \mathscr{W}$; (ii) $0 \leq \inf_{w,\beta}\Lambda(w;\beta), \sup_{w,\beta}\Lambda(w;\beta) \leq 1$; (iii) $\Lambda_x(\cdot;\beta)$ is continuously differentiable in β, and $\sup_{w,\beta}|\Lambda_x(w;\beta)| \leq c_2 < \infty$.

(d) There exists a symmetric, non-random matrix Ω, such that $||\Omega_n - \Omega|| = O_p(\delta_{\omega,n})$, where $\delta_{\omega,n} = o(1)$, and that $c_3^{-1} \le \lambda_{min}(\Omega) \le \lambda_{max}(\Omega) \le c_3$.

(e) There is a unique $\gamma_0 \in \Theta_\gamma$, and known function $L(\cdot)$ such that (i)

$$\ell(z) = \frac{1-Q_0}{Q_0} \cdot \frac{L(k(z)'\gamma_0)}{1-L(k(z)'\gamma_0)}.$$

(ii) $L(\cdot)$ is strictly increasing, twice continuously differentiable, with bounded first and second order derivatives; (ii) $\lim_{x\to-\infty} L(x) = 0$ and $\lim_{x\to\infty} L(x) = 1$; (iii) $0 < c_4 < L(k(z)'\gamma + t(z)'\lambda_j) \le c_5 < 1$ for all $(\gamma, \lambda_j) \in \Theta_\gamma \times \Theta_\lambda$, $j = s, a$, and $z \in \mathscr{Z}$.

(f) $\mathbb{E}[||j(Z)||^4] < +\infty$, where $j = k, t, e$.

Assumption 2.7(a) is standard in the microeconometric literature. Assumption 2.7(b) requires the conditional density $f_{Y|ZR}(\cdot|\cdot, 1)$ be bounded uniformly for all $(y, z) \in \mathscr{Y}_0\mathscr{Z}$. Assumption 2.7(c) imposes mild smoothness conditions on the parametric function $\Lambda(\cdot,\cdot;\cdot)$, requiring it to be bounded between the unit interval, thus behaving like a distribution function. Assumption 2.7(d) states that Ω_n is consistent for Ω, which is positive definite. Assumption 2.7(e) implies that the true "propensity score" is known up to finite dimensional γ_0. It also specifies smoothness and boundedness conditions on the parametric propensity score. Finally, due to the estimation of q_τ, we impose a finite fourth moment condition in Assumption 2.7(f), which is stronger than the usual square-integrability condition.

Lemma 2.2 Suppose that Assumptions 2.1–2.5 and Assumption 2.7 hold, then (i) $\widehat{\beta} \xrightarrow{p} \beta_0$; furthermore, (ii) suppose that the Jacobian matrix, M_Ω, as defined in (2.8.5), is invertible, then

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\beta(A_i; \theta_0, q_\tau) + o_p(1),$$

where $\psi_\beta(A; \theta_0, q_\tau)$ is given by (2.8.8), and (iii)

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma_\beta),$$

where $\Sigma_\beta \equiv \mathbb{E}[\psi_\beta(A; \theta_0, q_\tau)\psi_\beta(A; \theta_0, q_\tau)']$.

Lemma 2.2 shows that the parameters of $F_{Y_s|X_sZR=1}$ are consistently estimated by $\widehat{\beta}$. Furthermore, it admits an asymptotic linear representation with influence function given by $\psi_\beta(A; \theta_0, q_\tau)$, which plays a key role in establishing the large sample properties of UQE. Towards this ends, we need the following set of assumptions.

Assumption 2.8

(a) (i) $F_{Y|R=1}(\cdot)$ is absolutely continuous and differentiable over $y \in \mathscr{Y}$; (ii) $f_{Y|R=1}(\cdot)$ is uniformly continuous; (iii) the density $f_{Y|R=1}(y)$ is strictly bounded away from 0, three times continuously differentiable in y with uniformly bounded derivatives for y in $\mathscr{Y}_0$, such that $q_\tau \in \mathscr{Y}_0$.

(b) The kernel function $K_y(\cdot)$ is symmetric, continuous, bounded, with a compact support, and such that (i) $\int K_y(y)\, dy = 1$; (ii) $\int yK_y(y)dy = 0$.

(c) $b_y \to 0,\ \log(n)n^{-1}b_y^{-1} \to 0$, and $nb_y^5 \to c_6 < \infty$.

(d) (i) $\mathscr{X}$ is compact; (ii) G is continuously differentiable on $\mathscr{X}$ with strictly positive density.

Assumption 2.8(a) strengthens Assumption 2.5 and imposes stronger smoothness conditions on the distribution of Y_s. Assumption 2.8(b) states several regularity conditions on kernel functions, which is standard in the literature. Assumption 2.8(c) specifies admissible rate for the bandwidth parameter. We can choose $b_y = O(n_s^{-\kappa})$, for $\kappa \in [1/5, 1/2)$. Assumption 2.8(d) imposes support and smoothness conditions for the counterfactual target covariate.

Asymptotic properties of $\widehat{UQE}$ are formally characterized in the next theorem.

Theorem 2.3 Under Assumptions 2.1–2.5, 2.7, and 2.8, (i) the following linear expansions hold,

$$\widehat{UQE}_q(\tau, G) - UQE_q(\tau, G) = \frac{1}{n}\sum_{i=1}^{n} \psi_q + B_q(\tau, G, b_y) + o_p(n^{-1/2}).$$

Suppose in addition that Assumption 2.9 holds, (ii) then we have

$$\widehat{UQE}_p(\tau, G) - UQE_p(\tau, G) = \frac{1}{n}\sum_{i=1}^{n} \psi_p + B_p(\tau, G, b_y) + o_p(n^{-1/2}),$$

where, $\psi_j,\ j = p, q$, is defined in Section 2.7.2, $B_j(q_\tau, G, b_y) \equiv \frac{b_y^2 f''_{Y|R}(q_\tau|1)d_j(\theta_0,G)}{2f^2_{Y|R}(q_\tau|1)} \cdot \int y^2 K_y(y)dy$, and $d_j(\theta_0, G) \equiv \frac{1}{1-Q_0}\mathbb{E}[(1-R)\ell(Z)\Lambda_x(X, Z_1; \beta_0)g_j(X)]$, for $j = p, q$.

(iii) Therefore,

$$\sqrt{nb_y}(\widehat{UQE}_j(\tau, G) - UQE_j(\tau, G) - B_j(q_\tau, G, b_y)) \xrightarrow{d} N(0, \Sigma_j),$$

where, $\Sigma_j \equiv \frac{d_j^2(\theta_0,G)}{f^3_{Y|R}(q_\tau|1)Q_0} \int K_y^2(y)dy$, for $j = p, q$.

From the linear expansions in Theorem 2.3, we conclude that UQE converges at a rate that is slower than root-n. This result is mainly driven by the nonparametric estimation of the density $f_{Y|R=1}$, and therefore, the estimator is nonparametric in essence. Moreover, the asymptotic expansion includes an asymptotic bias term, $B(\tau, G, b_y)$. If we assume, as in Firpo et al. (2009a), $nb_y^5 \to 0$ or $\kappa < 1/5$, the bias vanishes asymptotically.

Remark 2.4 Estimators for the asymptotic variance of $UQE_p(\tau, G)$ and $UQE_q(\tau, G)$ can be constructed using their empirical counterparts. Specifically, let

$$\widehat{\Sigma}_j \equiv \frac{\widehat{d}_{j,n}(\widehat{\theta}, G)^2}{\widehat{f}^3_{Y|R}(\widehat{q}_\tau|1)\mathbb{E}_n[R]} \int K_y^2(y)dy,$$

where $\widehat{d}_{j,n}(\widehat{\theta}, G) \equiv \mathbb{E}_{n_a}[\widehat{\ell}(Z)\Lambda_x(W;\widehat{\beta})\widehat{g}_j(X)]$. Under a suitable rate condition on b_y, consistency of $\widehat{\Sigma}_j$ follows directly from the first two parts of Theorem 2.3. To achieve better finite-sample performance, we can add the root-n terms of the influence functions to the variance estimator, based on which, we propose the following improved variance estimator,

$$\widehat{\Sigma}_{j,imp} \equiv b_y \mathbb{E}_n[\widehat{\psi}_j(A;\widehat{\theta}, \widehat{q}_\tau, b_y)^2]. \tag{2.4.8}$$

In the above definition, $\widehat{\psi}_j(a;\widehat{\theta}, \widehat{q}_\tau, b_y)$ is a plug-in estimator of the influence function, $\psi_j(A;\theta_0, q_\tau, b_y)$, for $j = p, q$. A detailed description of the construction of $\widehat{\psi}$ can be found in Section 2.8.3.

Remark 2.5 Theorem 2.3 implies that tests of the unconditional quantile effect converges at a non-parametric rate in general. Nonetheless, for the null of zero, positive, and negative effects, we can still construct tests that have power against departures of the null at the parametric rate. For example, to test the null: $H_0 : UQE_j(\tau, G) = 0$, it is equivalent to test $H_0' : d_j(\beta_0, G) = 0$, as $UQE_j(\tau, G) = 0 \Leftrightarrow d_j(\beta_0, G) = 0$, for $j = p, q$. From Theorem 2.3, we know that $\widehat{d}_{j,n}(\widehat{\beta}, G)$ converges at the parametric rate. Moreover, we have

$$\sqrt{n}\widehat{V}_{d,j}^{-1/2}(\widehat{d}_{j,n}(\widehat{\theta}, G) - d_j(\theta_0, G)) \xrightarrow{d} N(0,1), \tag{2.4.9}$$

where $\widehat{V}_{d,j}$ is an estimator of $V_{d,j} \equiv \mathbb{E}[\psi_{d,j}(A;\theta_0, q_\tau)^2]$, with $\psi_{d,j}$, $j = p, q$, defined in Section 2.7.2. The result in (2.4.9) can be used to test H_0', applying standard testing procedures.

2.5 Empirical Illustration

We apply our identification and estimation methods to a variant of the Mincer's regression. Our main goal here is to demonstrate the bias from using potential instead of actual labor experience in human capital earnings models.

Identifying the causal relationship between earnings and human capital accumulation has been a focus of labor economic studies for decades. Traditionally, Mincer's regression has been widely used to quantify the link between labor wage, education and labor market experience.

Most datasets do not provide respondents' actual work histories. Therefore, many researchers choose to proxy the variable with potential work experience. The potential experience measure is usually calculated by subtracting years of schooling plus some constant (typically 6 years) from age. Despite the popularity of this practice, many labor economists believe that the return to actual experience tends to be biased when we employ the potential experience as proxy; see e.g. Regan and Oaxaca (2009). One of their main arguments is that any lapse in labor force participation would be implicitly assumed away when potential experience instead of the actual one is used. There is little reason to believe that the return to employed experience is the same as that of the unemployed period. Hence, it is still preferable to use the actual labor experience.

We use the 1970 wave of IPUMS as our main sample. The data is a 1-in-10,000 national random sample of the population. The outcome of interest is the natural log of yearly earnings. The target covariate, actual work experience, is missing from IPUMS. To apply the procedure described in Section 2.4, we need a dataset where the actual work experience is available. For that purpose, we use the 1972 wave of PSID as cleaned by Hirukawa et al. (2020). Detailed work histories are available in PSID. Therefore, it allows us to recover the actual labor market experience. However, running analysis directly with PSID may not be ideal due to the fact that it is not nationally representative. Our method is able to address this issue by combining information from both samples.

To estimate $F_{Y_s|X_sZ_1R=1}$, we consider the following specification,

$$\begin{aligned}\mathbb{P}(log(Income) \leq y) = \Lambda(&\beta_0 + \beta_1 educ + \beta_2 black + \beta_3 south \\ &+ \beta_4 married + \beta_5 exper_r + \beta_6 exper_r^2),\end{aligned} \tag{2.5.1}$$

where $exper_r$ stands for individual's actual or realized work experience, $educ$ denotes the highest grade completed by the respondent, $black, married$, and $south$ are dummy variables which take one if the person is black, married, and lives in the south, respectively. The actual work experience serves as our X_s. It enters (2.5.1) with linear and quadratic terms. We let $(educ, black, south, married)$ be the set of included instruments Z_1, and the potential experience, $exper_p$, be the excluded instrument Z_2.

Mincer et al. (1974) derives the relationship between schooling, labor market experience and earnings by means of an accounting identity model. We assume that realized labor experience, rather than potential experience as constructed by econometricians, determines post-school investment, and therefore, observed earnings. This belief is embodied in Assumption 2.2, which requires that earnings are independent with potential experience, conditional on actual experience and education. This restriction holds if the mechanism that governs the discrepancy between potential and actual experience is unrelated to the wage determination process. Note that it in principle rules out cases where

Table 2.1: Summary statistics

Variable	Mean	St. Dev.	Min	25th Pctl.	Median	75th Pctl.	Max
Data Source A: The IPUMS Sample (1970s)							
Income	7,923.98	6,218.37	50	4,050	7,050	10,050	50,000
Log(Income)	8.64	0.97	3.91	8.31	8.86	9.22	10.82
Age	38.29	13.88	17	26	37	50	65
Education	11.43	2.71	5	10	12	13	17
Black	0.07	0.25	0	0	0	0	1
South	0.26	0.44	0	0	0	1	1
Married	0.75	0.43	0	0	1	1	1
Potential Experience	20.86	14.7	0	8	20	33	53
Data Source B: The PSID Sample							
Income	8,966.52	5,905.47	50	5,069	8,000	11,359	70,000
Log(Income)	8.88	0.74	3.91	8.53	8.99	9.34	11.16
Age	37.8	12.37	17	26	37	47	65
Education	12.14	3.06	5	11	12	16	17
Black	0.27	0.44	0	0	0	1	1
South	0.41	0.49	0	0	0	1	1
Married	0.89	0.31	0	1	1	1	1
Potential Experience	19.65	13.42	0	7	18	30	53
Actual Experience	18.87	12.23	0	8	18	28	56
Data Source C: The IPUMS Sample (1980s)							
Potential Experience	18.99	14.49	0	6	16	31	53

Notes: Summary statistics for IPUMS and PSID. The top panel uses male subsample (aged between 17 to 65) from the 1970 wave of IPUMS with a sample size of 5,807. The middle panel uses the male subsample (aged between 17 to 65) from the 1972 wave of PSID with a sample size of 2,339. The bottom panel uses male subsample (aged between 17 to 65) from the 1980 wave of IPUMS with a sample size of 533,517.

individuals leave labor market in response to wage rate fluctuations.

Visual check of the actual-experience-specific age-income profiles can serve as a preliminary test of the exclusion restriction. There are many reasons why such a parsimonious model is often refuted by data. See Heckman et al. (2006), Lemieux (2006), and the references therein for a detailed discussion of the empirics. Nevertheless, we believe that our modification of the benchmark Mincer regression suffices for an illustrative purpose.

Due to the relatively large support, we treat experience as a continuous variable. We assume that, given potential experience and the set of controls, actual labor experience follows the same distribution in the two samples, which implies that Assumption 2.1(e) holds. Additionally, we assume the errors of structural equations follow a joint normal distribution. Therefore, Assumption 2.3 follows by Example 2.1.

We provide estimation results when $\Lambda(\cdot)$ takes either the logistic link or the probit link. To implement the AST estimator, we choose $j(Z) = (Z_1', Z_2')'$, for $j = k,t,e$. The density of Y_s is estimated using $b_y = n_s^{-0.01} b_{n,0}$, where $b_{n,0} \equiv 1.06 \min\{\sigma(Y_s), interquartile(Y_s)\} n_1^{-0.2}$ is the usual "rule-of-thumb" bandwidth.

Table 2.1 reports the descriptive statistics for the two samples. Following the standard practice in labor economics,

Table 2.2: Estimation results for unconditional quantile effects

Quantile Level	0.25	0.5	0.75	0.25	0.5	0.75
	Logit Link			Probit Link		
MDS						
$UQE_{2s}(\tau)$	-0.2297	-0.0612	-0.0296	-0.2304	-0.0619	-0.0299
	(0.0587)	(0.0125)	(0.0054)	(0.0576)	(0.0124)	(0.0055)
$H_0 : UQE_{2s} = 0$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MQS						
$UQE_{2s}(\tau)$	-0.2159	-0.0611	-0.0305	-0.2159	-0.0615	-0.0308
	(0.0779)	(0.0164)	(0.0063)	(0.0778)	(0.0164)	(0.0063)
$H_0 : UQE_{2s} = 0$	0.0043	0.0002	0.0000	0.0043	0.0002	0.0000
MLS						
$UQE_{2s}(\tau)$	0.0956	0.0241	0.0135	0.0957	0.0243	0.0143
	(0.0154)	(0.0020)	(0.0011)	(0.0151)	(0.0020)	(0.0011)
$H_0 : UQE_{2s} = 0$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

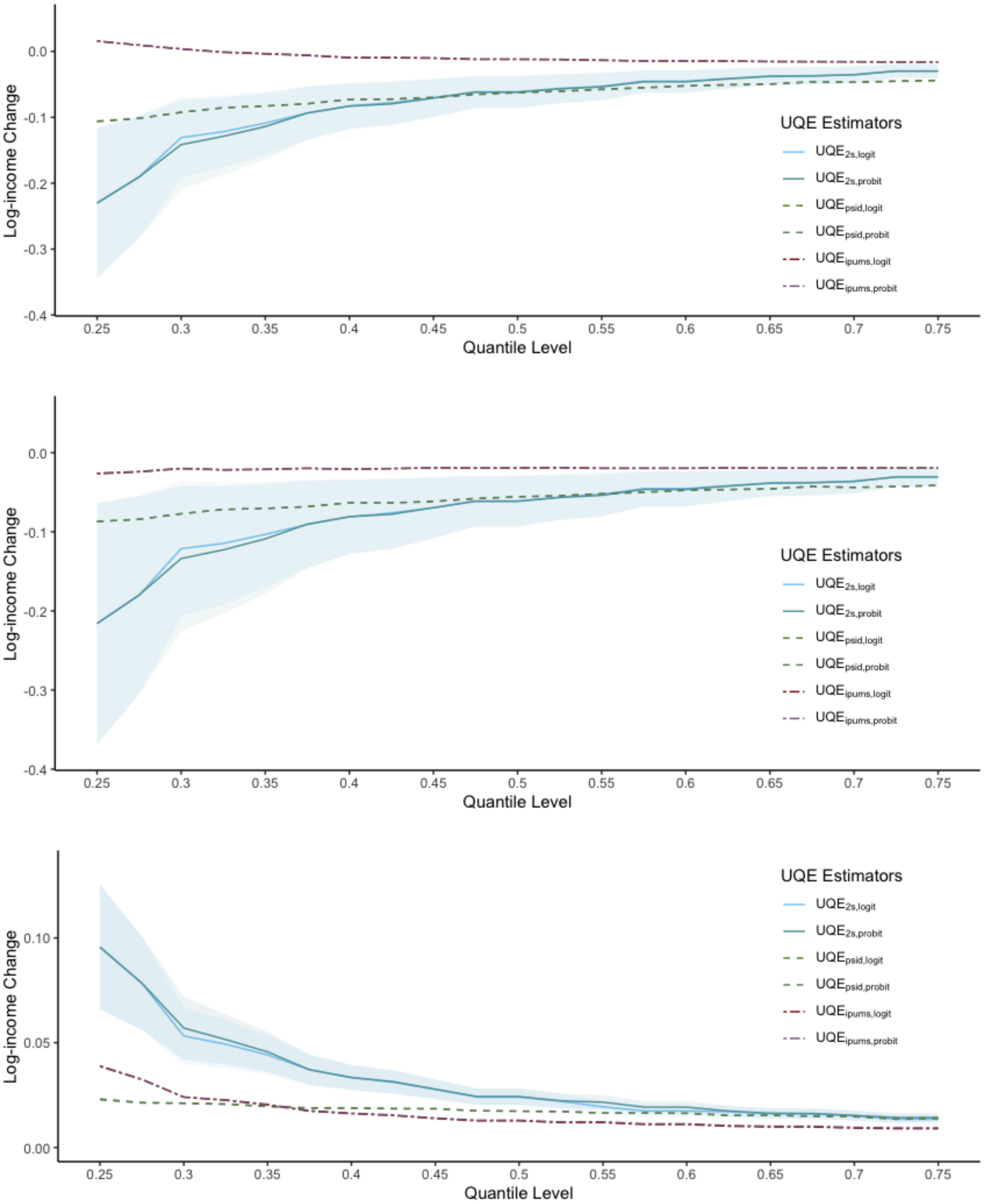
Notes: In each panel, the first two rows report point estimates and standard error using our two sample estimator. The last row of each panel reports the p-value associated with the Wald test of zero effect.

we use only the data on men aged between 17 and 65 when the surveys are taken. To ensure that Assumption 2.1(a) holds, we further trim the IPUMS sample to match the sample bounds observed in the PSID data set. This leaves us with a sample of $n_s = 5,807$ respondents for IPUMS and $n_a = 2,339$ for PSID.There are considerable differences between the two datasets. Individuals who are black, married and/or lives in the south are over-represented in PSID compared to the nationally representative IPUMS. On average, an individual in PSID has 0.78 years more potential experience than actual experience.

For UQE with MDS and MQS, we take the smoothed empirical distribution of $exper_p$ from the 1980 wave of IPUMS 1-in-100 sample (trimmed to match the support of the PSID sample) as the target counterfactual distribution. The policy question we would like to answer with this counterfactual is as follows: What is the unconditional quantile effect if the distribution of labor market experience shifts marginally towards that is observed in the 1980s. Due to the large sample size of the counterfactual sample ($n = 533,517$), we can ignore the sampling variation and treat the target distribution as known. As shown in Table 2.1, we find that less-experienced workers tend to have even fewer years of experience in the counterfactual scenario than in the status quo, and the opposite is true for workers closer to the right tail of the distribution.

We report estimation results in Table 2.2 and Figure 2.2. A few remarks are in order. First, our estimates suggest that the counterfactual effect of a marginal shift in the distribution of actual experience is heterogeneous across income groups. The effect is larger in magnitude for the lower-income groups as expected. When MDS and MQS are considered, the quantile effects are uniformly negative and the shapes of the effect curves are similar. The marginal shift could decrease the (log) earning by anything between 0.03 and 0.23 across income quantiles. For reference,

Figure 2.2: Unconditional quantile effect of actual experience on log(Earnings).



Notes: The top panel: Results for UQE with MDS. The middle panel: Results for UQE with MQS. The bottom panel: Results for UQE with MLS. All three plots contain the UQE of potential experience based on IPUMS (two-dash lines), the UQE of actual experience based on PSID (dashed lines), the two-sample UQE (solid lines), and the two-sided 95% confidence intervals based on the improved variance estimator (shaded area).

when the logistic link is assumed, the marginal effect of MDS amounts to a reduction of 20.5% in annual earnings for individuals the first quartile, 5.9% at the median, and 2.9% at the third quartile, respectively. The MLS estimates are of different signs from MDS and MQS. The marginal upward shift in the actual experience would increase a median worker's income by 2.4%.

Next, we consider the bias caused by using potential experience in lieu of the actual experience. We note that, one-sample UQE estimates based on IPUMS tend to be smaller in magnitude at lower income quantiles than that based on the combined data. Eventually, the two estimators converge at higher income levels.[6]

2.6 Concluding Remarks

In this paper, we propose a framework to identify and estimate unconditional quantile policy effect under data combination. We establish the identification of UQE under two main conditions: a rank similarity assumption and a conditional independence assumption, based on which, we provide estimators for the identified UQE and derive their large sample properties.

Our current approach can be extended in the following directions. First, although we have restricted our attention to the quantile effect throughout this paper, our results can be easily extended to other statistical functionals such as mean, interquartile, and inequality measures. It would be interesting to see how the identification requirements change with respect to the functional we adopt. Second, we have focused exclusively on the pointwise identification and inference. While extension to uniform results seem straightforward, it comes at a cost of stronger cross-sample restrictions. Under such assumptions, conditional quantile regression is likely feasible. Comparing conditional and unconditional quantile effects, as in Firpo et al. (2009b), under our two-sample structure, would also be an interesting direction for future research.

2.7 Appendix

2.7.1 Proofs of Lemmas and Theorems in Section 2.3

Proof of Lemma 2.1: We provide proof for the nonparametric identification here. The proof for the parametric case follows along exactly the same line and is omitted. We shall show (2.3.1) first,

$$\begin{aligned}
\mathbb{E}[1(Y \leq q_\tau)|Z, R=1] &= \mathbb{E}[\mathbb{E}[1(Y_s \leq q_\tau)|X_s, Z, R=1]|Z, R=1] \\
&= \mathbb{E}[\mathbb{E}[1(Y_s \leq q_\tau)|X_s, Z_1, R=1]|Z, R=1] \\
&= \mathbb{E}[\Lambda(X_s, Z_1)|Z, R=1] \\
&= \mathbb{E}[\Lambda(X_a, Z_1)|Z, R=0] \\
&= \mathbb{E}[\Lambda(W)|Z, R=0],
\end{aligned}$$

[6] Our analysis is local to the direction of counterfactual change, and therefore, does not allow the result to be extrapolated globally. For the same reason, the comparison between UQE_{2s} and UQE_{psid} is not meaningful.

where the second equality is by Assumption 2.2, and the fourth line follows by Assumption 2.1(e). Likewise for (2.3.2),

$$\begin{aligned}
\mathbb{E}[R1(Y \leq q_\tau)|Z] &= \mathbb{E}[1(Y \leq q_\tau)|Z, R=1] \cdot \mathbb{P}[R=1|Z] \\
&= \mathbb{E}[\Lambda(W)|Z, R=0] \cdot r(Z) \\
&= \mathbb{E}[(1-R)\Lambda(W)|Z] \cdot \frac{r(Z)}{1-r(Z)}.
\end{aligned}$$

Thus, Lemma 2.1 follows immediately from (2.3.1) (or (2.3.2)) and Assumption 2.3. ■

Lemma 2.3 Suppose (i) $\Lambda(w;\beta)$ is measurable with respect to w for all $\beta \in \Theta_\beta$; (ii) W is bounded complete for Z, relative to the auxiliary population; (iii) $\Lambda(w;\beta)$ is differentiable with respect to β; and (iv) $\partial\Lambda(\cdot;\beta)/\partial\beta$ is uniformly bounded and $\partial\Lambda(\cdot;\beta)/\partial\beta \not\equiv 0$ for all $\beta \in \Theta_\beta$. Then, under Assumptions 2.1 and 2.2, β_0 can be uniquely identified from (2.3.1) or (2.3.2).

Proof of Lemma 2.3: From Lemma 2.1, we know that β_0 solves (2.3.1) or (2.3.2). It remains to show uniqueness. Suppose, there is β_1, $\beta_1 \neq \beta_0$, that solves (2.3.1), then, $\mathbb{E}[\Lambda(W;\beta_1) - \Lambda(W;\beta_0)|Z, R=0] = 0$. By MVT, this and (iii) implies that $\mathbb{E}[\partial\Lambda(W;\beta)/\partial\beta|_{\beta=\widetilde{\beta}}|Z, R=0](\beta_1 - \beta_0) = 0$, for some value between β_0 and β_1. Condition (i), (ii), and (iv) then imply that $\mathbb{E}[\partial\Lambda(W;\beta)/\partial\beta|_{\beta=\widetilde{\beta}}|Z, R=0] \not\equiv 0$, which leads to a contradiction. ■

Proof of Theorem 2.1: We shall first prove the identification result for a fixed counterfactual distribution. Next, we take the derivative of the counterfactual experiments with respect to t. The result of Theorem 2.1 then follows by the fact that Hadamard derivative operator of the quantile functional is linear. For any $t \leq t_0$, fix $\varepsilon_t \in \Phi^*$, and we have that

$$\begin{aligned}
&F_{\widetilde{Y}_{s,t}|R}(q_\tau|1) \\
&= \int P(g_s(\widetilde{X}_{s,t}, \widetilde{Z}_{1,t}, \widetilde{\varepsilon}_{s,t}) \leq q_\tau | \widetilde{X}_{s,t} = x, \widetilde{Z}_{1,t} = z_1, \widetilde{R}_t = 1) dF_{\widetilde{X}_{s,t}\widetilde{Z}_{1,t}|\widetilde{R}_t}(x, z_1|1) \\
&= \int P(g_s(G_t^{-1}(\widetilde{U}_{s,t}), \widetilde{Z}_{1,t}, \widetilde{\varepsilon}_{s,t}) \leq q_\tau | \widetilde{U}_{s,t} = u, \widetilde{Z}_{1,t} = z_1, \widetilde{R}_t = 1) dF_{\widetilde{U}_{s,t}\widetilde{Z}_{1,t}|\widetilde{R}_t}(u, z_1|1) \\
&= \int P(g_s(G_t^{-1}(u), Z_1, \varepsilon_s) \leq q_\tau | Z_1 = z_1, R = 1) dF_{U_s Z_1|R}(u, z_1|1) \\
&= \int P(g_s(G_t^{-1}(u), Z_1, \varepsilon_s) \leq q_\tau | U_s = u, Z_1 = z_1, R = 1) dF_{U_s Z_1|R}(u, z_1|1) \\
&= \int P(g_s(X_s, Z_1, \varepsilon_s) \leq q_\tau | X_s = G_t^{-1}(u), Z_1 = z_1, R = 1) dF_{U_s Z_1|R}(u, z_1|1) \\
&= \int P(g_s(X_s, Z_1, \varepsilon_s) \leq q_\tau | X_s = G_t^{-1}(u), Z_1 = z_1, R = 1) dF_{U_s Z|R}(u, z|1) \\
&= \int P(g_s(X_s, Z_1, \varepsilon_s) \leq q_\tau | X_s = G_t^{-1}(F_{X_s}(x)), Z_1 = z_1, R = 1) dF_{X_s Z|R}(x, z|1)
\end{aligned}$$

$$
\begin{aligned}
&= \int F_{Y_s|X_sZ_1R}(q_\tau|G_t^{-1}(F_{X_s|R}(x|1)),Z_1,1)dF_{X_sZ|R}(x,z|1)\\
&= \int F_{Y_s|X_sZ_1R}(q_\tau|G_t^{-1}(F_{X_s|R}(x|1)),Z_1,1)\frac{r(z)(1-Q_0)}{Q_0(1-r(z))}dF_{XZ|R}(x,z|0)\\
&= \mathbb{E}\left[F_{Y_s|X_sZ_1R}(q_\tau|G_t^{-1}(F_{X_s|R}(X|1)),Z_1,1)\frac{r(Z)(1-Q_0)}{Q_0(1-r(Z))}|R=0\right]\\
&= \frac{1}{1-Q_0}\mathbb{E}\left[(1-R)\ell(Z)\cdot F_{Y_s|X_sZ_1R}(q_\tau|G_t^{-1}(F_{X|R}(X|1)),Z_1,1)\right],
\end{aligned}
$$

where the second line follows by the definition of $F_{\widetilde{Y}_{s,t}}(q_\tau)$, the third one comes from the definition of $\widetilde{U}_{s,t}$ and a change of variable from x to u, the fourth equality follows by the construction of Φ^* and Assumptions 2.4(a) and (b), the fifth line is again by Assumption 2.4(a), the eighth one follows by the definition of U_s and standard change-of-variable argument, the tenth line is by Assumptions 2.1(a)–(c) and Bayes' Law.

To obtain the marginal distributional effect, we take derivative of $F^{G_t}_{\widetilde{Y}_s|R}(q_\tau|1)$ with respect to t and evaluate it at $t=0$. For the marginal distributional shift,

$$
\begin{aligned}
\left.\frac{\partial F_{\widetilde{Y}_{s,t}|R}(q_\tau|1)}{\partial t}\right|_{t=0} =& \int \frac{\partial F_{Y_s|X_sZ_1R}(q_\tau|x,z_1,1)}{\partial x}\cdot\left.\frac{\partial G_{t,p}^{-1}(F_{X_s|R}(x|1))}{\partial t}\right|_{t=0}\\
&\qquad\qquad\qquad\qquad\cdot\frac{r(z)(1-Q_0)}{Q_0(1-r(z))}dF_{W|R}(w|0)\\
=&\frac{1}{1-Q_0}\mathbb{E}\left[(1-R)\ell(Z)\cdot\frac{\partial\Lambda(X,Z_1)}{\partial x}\cdot\frac{\partial G_{t,p}^{-1}(F_{X_s|R}(x|1))}{\partial t}|_{t=0}\right].
\end{aligned}
$$

Observe that $\frac{\partial G_{t,p}^{-1}(\cdot)}{\partial t}|_{t=0}$ is the pathwise derivative of the inverse map $H\mapsto H^{-1}$ at $F_{X_s|R=1}$ in the direction of $G-F_{X_s|R=1}$. By Lemma 3.9.23 in Van Der Vaart and Wellner (1996), the inverse map is Hadamard differentiable under the conditions specified in the theorem, with the derivative map given by,

$$
\phi\mapsto-(\phi/h)\circ H^{-1},
$$

where h is the first-order derivative of H. Let $\phi(\cdot)=G(\cdot)-F_{X_s|R=1}(\cdot)$ and $H=F_{X_s|R=1}$, it follows immediately that, for all $u\in[0,1]$,

$$
\left.\frac{\partial G_{t,p}^{-1}(u)}{\partial t}\right|_{t=0} = -\frac{G(F^{-1}_{X_s|R=1}(u))-u}{f_{X_s|R=1}(F^{-1}_{X_s|R=1}(u))},
$$

and hence, for all $x\in\mathscr{X}$,

$$
\left.\frac{\partial G_{t,p}^{-1}(F_{X_s|R}(x|1))}{\partial t}\right|_{t=0} = \frac{F_{X_s|R}(x|1)-G(x)}{f_{X_s|R=1}(x)}.
$$

Analogously, for marginal quantile shift,

$$\left.\frac{\partial F_{\widetilde{Y}_{s,t}|R}(q_\tau|1)}{\partial t}\right|_{t=0} = \int \frac{\partial F_{Y_s|X_sZ_1R}(q_\tau|x,z_1,1)}{\partial x} \cdot \left.\frac{\partial G_{t,q}^{-1}(F_{X_s|R}(x|1))}{\partial t}\right|_{t=0} \cdot \frac{r(z)(1-Q_0)}{Q_0(1-r(z))} dF_{W|R}(w|0)$$
$$= \frac{1}{1-Q_0}\mathbb{E}\left[(1-R)\ell(Z)\cdot\frac{\partial\Lambda(X,Z_1)}{\partial x}\cdot(G^{-1}(F_{X|R=1}(X))-X)\right],$$

where the second equality follows from Lemma 2.1 and the definition of $G_{t,q}^{-1}(\cdot)$.

To identify $F_{X_s|R=1}$, we exploit the following fact

$$\begin{aligned}
F_{X_s|R=1}(\cdot) &= \mathbb{E}[1(X \leq \cdot)|R=1] \\
&= \int_{\mathscr{Z}}\int_{\mathscr{X}} 1(x \leq \cdot) dF_{X|ZR}(x|z,1)dF_{Z|R}(z|1) \\
&= \int_{\mathscr{Z}}\int_{\mathscr{X}} 1(x \leq \cdot)\cdot\frac{(1-Q_0)r(z)}{Q_0(1-r(z))} dF_{X|ZR}(x|z,0)dF_{Z|R}(z|0) \\
&= \mathbb{E}\left[\frac{(1-Q_0)r(Z)}{Q_0(1-r(Z))}1(X \leq \cdot)|R=0\right] \\
&= \frac{1}{1-Q_0}\mathbb{E}\left[(1-R)\ell(Z)1(X \leq \cdot)\right],
\end{aligned}$$

where the third line is due to Assumption 2.1 and Bayes' Law.

Theorem 2.1 then follows from Assumption 2.5, $q_\tau = F_{Y_s|R=1}^{-1}(\tau)$, and the fact that the Hadamard derivative of the quantile functional is $\nu_\tau'(\phi) = -\frac{\phi}{f_{Y_s|R=1}} \circ F_{Y_s|R=1}^{-1}(\tau)$, which is linear in ϕ. ■

Proof of Theorem 2.2: First, we fix $U_s \in \mathscr{U}_s$ and $\phi_t \in \Phi^*$, for $t \leq t_0$. By construction, there exists $\widetilde{U}_{s,t} \in \widetilde{\mathscr{U}_{s,t}}$ such that $(\widetilde{U}_{s,t}|Z_1, R=1) \overset{d}{=} (U_s|Z_1,R=1)$. Now we rewrite $F_{\widetilde{Y}_{s,t}|R=1} \in \mathscr{F}_{\widetilde{Y}_{s,t}|R=1}$ in terms of U_s and Z_1. Let $x^0 = -\infty$, and we have that

$$\begin{aligned}
&F_{\widetilde{Y}_{s,t}|R=1}(q_\tau) \\
&= \int P(g_s(\widetilde{X}_{s,t},\widetilde{Z}_{1,t},\widetilde{\varepsilon}_{s,t}) \leq q_\tau|\widetilde{X}_t = x, \widetilde{Z}_{1,t} = z_1, \widetilde{R}_t = 1)dF_{\widetilde{X}_{s,t}\widetilde{Z}_{1,t}|\widetilde{R}_t}(x,z_1|1) \\
&= \sum_{j=1}^{l}\int P(g_s(\widetilde{X}_{s,t},\widetilde{Z}_{1,t},\widetilde{\varepsilon}_{s,t}) \leq q_\tau|\widetilde{X}_{s,t} = x^j, \widetilde{Z}_{1,t} = z_1, \widetilde{R}_t = 1) \\
&\qquad \cdot P(\widetilde{U}_{s,t} \in (G_{t,p}(x^{j-1}), G_{t,p}(x^j)]|\widetilde{Z}_{1,t} = z_1, \widetilde{R}_t = 1)dF_{\widetilde{Z}_{1,t}|\widetilde{R}_t}(z_1|1) \\
&= \sum_{j=1}^{l}\int P(g_s(X_s,Z_1,\varepsilon_s) \leq q_\tau|X = x^j, Z_1 = z_1, R = 1) \\
&\qquad \cdot P(U_s \in (G_{t,p}(x^{j-1}), G_{t,p}(x^j)]|Z_1 = z_1, R = 1)dF_{Z_1|R}(z_1|1)
\end{aligned}$$

$$
\begin{aligned}
=&\sum_{j=1}^{l}\int F_{Y_s|X_sZ_1R}(q_\tau|x^j,z_1,1)\cdot(P(U_s\in(F_{X_s|R}(x^{j-1}|1),F_{X_s|R}(x^j|1)]|Z_1=z_1,R=1)\\
&+P(U_s\in(G_{t,p}(x^{j-1}),G_{t,p}(x^j)]|Z_1=z_1,R=1)\\
&-P(U_s\in(F_{X_s|R}(x^{j-1}|1),F_{X_s|R}(x^j|1)]|Z_1=z_1,R=1))dF_{Z_1|R}(z_1|1)\\
=&F_{Y_s|R}(q_\tau|1)-\sum_{j=1}^{l}\int\Lambda(x^j,z_1)(P(U_s\in(G_{t,p}(x^{j-1}),G_{t,p}(x^j)]|Z_1=z_1,R=1)\\
&-P(U_s\in(F_{X_s|R}(x^{j-1}|1),F_{X_s|R}(x^j|1)]|Z_1=z_1,R=1))dF_{Z_1|R}(z_1|1). \qquad (2.7.1)
\end{aligned}
$$

For the second term on the right hand side of the last equality, we have

$$
\begin{aligned}
&P(U_s\in(G_{t,p}(x^{j-1}),G_{t,p}(x^j)]\mid Z_1=z_1,R=1)\\
&\quad-P(U_s\in(F_{X_s|R}(x^{j-1}|1),F_{X_s|R}(x^j|1)]\mid Z_1=z_1,R=1))\\
&=(F_{U_s|Z_1R}(G_{t,p}(x^j)\mid z_1,1)-F_{U_s|Z_1R}(F_{X_s|R}(x^j|1)\mid z_1,1)\\
&\quad-(F_{U_s|Z_1R}(G_{t,p}(x^{j-1})\mid z_1,1))-F_{U_s|Z_1R}(F_{X_s|R}(x^{j-1}|1)\mid z_1,1))\\
&=(G_{t,p}(x^j)-F_{X_s|R}(x^j|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j,t}\mid z_1,1)\\
&\quad-(G_{t,p}(x^{j-1})-F_{X_s|R}(x^{j-1}|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j-1,t}\mid z_1,1)\\
&=t\cdot(G(x^j)-F_{X_s|R}(x^j|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j,t}\mid z_1,1)\\
&\quad-t\cdot(G(x^{j-1})-F_{X_s|R}(x^{j-1}|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j-1,t}\mid z_1,1),
\end{aligned}
$$

where $\widetilde{u}_{j,t}$ is some value between $G_{t,p}(x^j)$ and $F_{X_s|R}(x^j|1)$, and is potentially dependent on z_1. The last equality is due to MVT. Using the above result, (2.7.1) becomes

$$
\begin{aligned}
&F_{Y_s|R}(q_\tau|1)-\sum_{j=1}^{l}t\cdot\int\Lambda(x^j,z_1)\cdot((G(x^j)-F_{X_s|R}(x^j|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j,t}\mid z_1,1)\\
&\qquad-(G(x^{j-1})-F_{X_s|R}(x^{j-1}|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j-1,t}\mid z_1,1))dF_{Z_1|R}(z_1|1)\\
=&F_{Y_s|R}(q_\tau|1)-\sum_{j=2}^{l}t\cdot\int(\Lambda(x^{j-1},z_1)-\Lambda(x^j,z_1))\\
&\qquad\cdot(G(x^{j-1})-F_{X_s|R}(x^{j-1}|1))\cdot f_{U_s|Z_1R}(\widetilde{u}_{j-1,t}\mid z_1,1)dF_{Z_1|R}(z_1|1),
\end{aligned}
$$

where the equality follows by rearranging terms, the fact that $G(x^0)=F_{X_s|R}(x^0|1)=0$, and that $G(x^l)=F_{X_s|R}(x^l|1)=1$.

The pathwise derivative can thus be calculated as

$$\begin{aligned}
&\lim_{t\downarrow 0}\frac{F_{\widetilde{Y}_{s,t}|R=1}(q_\tau)-F_{Y_s|R}(q_\tau|1)}{t}\\
&=\sum_{j=2}^{l}\int(\Lambda(x^{j-1},z_1)-\Lambda(x^{j},z_1))\\
&\qquad\cdot(G(x^{j-1})-F_{X_s|R}(x^{j-1}|1))dF_{Z_1|U_sR}(z_1\mid F_{X_s|R}(x^{j-1}|1),1),
\end{aligned}$$

where the second line is due to the dominated convergence theorem, Bayes's Law, and the fact that $U_s|R=1$ follows the standard uniform distribution. Therefore, by Lemma 2.1 and the linearity of $\nu'_\tau(\cdot)$,

$$\begin{aligned}
UQE_p(\tau,G)\in\Bigg[&\inf_{U_s\in\mathscr{U}_s}\sum_{j=2}^{l}\int h_{q_\tau}(x^j,x^{j-1},z_1)dF_{Z_1|U_sR}(z_1\mid F_{X_s|R}(x^{j-1}|1),1),\\
&\sup_{U_s\in\mathscr{U}_s}\sum_{j=2}^{l}\int h_{q_\tau}(x^j,x^{j-1},z_1)dF_{Z_1|U_sR}(z_1\mid F_{X_s|R}(x^{j-1}|1),1)\Bigg].
\end{aligned}$$

Using a similar argument as in the proof of Theorem 5 in Rothe (2012), we can show that for $j=1,\dots,l$, $\{F_{Z_1|U_sR}(z_1|U_s=F_{X_s|R}(x^j|1),R=1):U_s\in\mathscr{U}_s\}$ is the set of all multivariate distribution functions with support equal to $\text{Supp}(F_{Z_1|R=1})$. To see this, note that for $j=1,\dots,l$, $F_{Z_1|U_sR}(\cdot|U_s=F_{X_s|R}(x^j|1),R=1)=C_1^{U_s}(F_{X_s|R}(x^j|1),F_{Z_1|R}(\cdot|1))$, where the conditional copula, C^{U_s}, is defined by $C^{U_s}(F_{U_s|R}(u|1),F_{Z_1|R}(z_1|1))\equiv F_{U_sZ_1|R}(u,z_1|1)$, and $C_1^{U_s}$ is the partial derivative of C^{U_s} with respect to the first argument. By the construction of Φ, the set of $C^{U_s}(\cdot,\cdot)$ for $U_s\in\mathscr{U}_s$ is equivalent to the identified set of the conditional copula of X_s and Z_1 given $R=1$, $C^{X_s}(\cdot,\cdot)$, where $C^{X_s}(F_{X_s|R}(x|1),F_{Z_1|R}(z_1|1))$ $\equiv F_{X_sZ_1|R}(x,z|1)$, for all $x\in\{x^1,\dots,x^l\}$. Then, the desired result follows by applying an extension of Theorem 2.2.7 in Nelsen (2007).

Without loss of generality, we focus on the upper bound for now. By appropriately choosing Dirac measures with unit masses on $\{z_j^*\}_{j\in\mathscr{J}_+}$ and $\{z_j^\dagger\}_{j\in\mathscr{J}_-}$, It is straightforward to show that,

$$\begin{aligned}
&\sup_{U_s\in\mathscr{U}_s}\sum_{j=2}^{l}\int h_{q_\tau}(x^j,x^{j-1},z_1)dF_{Z_1|U_sR}(z_1\mid F_{X_s|R}(x^{j-1}|1),1)\\
&\quad=\sum_{j\in\mathscr{J}_+}h_{q_\tau}(x^j,x^{j-1},z_{1,j}^*)+\sum_{j\in\mathscr{J}_-}h_{q_\tau}(x^j,x^{j-1},z_{1,j}^\dagger).
\end{aligned}\tag{2.7.2}$$

The right hand side of (2.7.2) is identified under the support condition in Assumption 2.1(a). The proof for the lower bound follows by an analogous argument. ■

2.7.2 Asymptotic Linear Representation of UQE Estimators

We specify additional regularity conditions in Theorem 2.3 and provide linear expansions for $\widehat{UQE}_p$ and $\widehat{UQE}_q$ in this section, the proofs of which are contained in Section 2.8.

Assumption 2.9

(a) $f_{X|ZR=0}$ is uniformly bounded, twice continuously differentiable with uniformly bounded first and second order derivatives on $\mathscr{X} \times \mathscr{Z}$.

(b) (i) $K_x(\cdot)$ is a second order symmetric kernel function; (ii) the support of K_x is continuous, bounded, with compact support, $K_x(\cdot)$, and such that $\int K_x(x)dx = 1$, $\int xK_x(x)dx = 0, \int x^2 K_x(x)dx > 0$, and $\int K_x^2(x)dx < \infty$.

(c) (i) $nb_x/log(n) \to \infty$ and (ii) $nb_x^4 \to 0$.

For $j = p, q$, the asymptotic linear representation of $\widehat{UQE}_j$ is given as follows:

$$\begin{aligned}
&\widehat{UQE}_j(\tau,G) - UQE_j(\tau,G) - B_j(\tau,d,b_y) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_{f_y,j}(A_i;\theta_0,q_\tau,G) - \frac{1}{f_{Y|R}(q_\tau|1)}\psi_{d,j}(A_i;\theta_0,q_\tau,G)\right\} + o_p(n^{-1/2}b_y^{-1/2} + b_y^2), \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi_j(A_i;\theta_0,q_\tau,b_y) + o_p(n^{-1/2}b_y^{-1/2} + b_y^2), \qquad (2.7.3)
\end{aligned}$$

where

$$\begin{aligned}
\psi_{f_y,j}(a;\theta_0,q_\tau,G) \equiv& \frac{d_j(\theta_0,G)}{f^2_{Y|R}(q_\tau|1)}\frac{r}{Q_0}\Big(K_{b_y}(y-q_\tau) \\
&- \mathbb{E}[K_{b_y}(Y-q_\tau)|R=1] - \frac{(1(y \le q_\tau) - \tau)f'_{Y|R}(q_\tau|1)}{f_{Y|R}(q_\tau|1)}\Bigg), \qquad (2.7.4)
\end{aligned}$$

$$\begin{aligned}
\psi_{d,j}(a;\theta_0,q_\tau,G) \equiv& \big(M_{\theta,j}(\theta_0)'\psi_\theta(a;\theta_0,q_\tau) \\
&+ \psi_{g,j}(a;\theta_0,G) + \frac{(1-r)\ell(z)\Lambda_x(w;\beta_0)g_j(x)}{1-Q_0} - \frac{rd_j(\theta_0,G)}{Q_0}\Bigg). \qquad (2.7.5)
\end{aligned}$$

In the above equation, $\psi_\theta(a;\theta_0,q_\tau)$ is given in (2.8.6),

$$M_{\theta,j}(\theta_0) \equiv \mathbb{E}\left[\frac{1-R}{Q_0}\begin{pmatrix}\Lambda_x(W;\beta_0)\big(\nabla_{L,\theta_0}(Z)\cdot g_j(X) + G_{j,\theta_0}(X)\big) \\ \dfrac{L_0(Z)}{1-L_0(Z)}\Lambda_{x,\beta}(W;\beta_0)g_j(X)\end{pmatrix}\right], \qquad (2.7.6)$$

$$\nabla_{L,\theta}(z) \equiv \begin{pmatrix} \dfrac{L_s'(z)(1-L_a(z))+L_s(z)L_a'(z)}{(1-L_a(z))^2}\cdot k(z) \\ \dfrac{L_s'(Z)}{1-L_a(Z)}\cdot t(Z) \\ \dfrac{L_s(Z)L_a'(Z)}{(1-L_a(Z))^2}\cdot t(Z) \end{pmatrix}, \tag{2.7.7}$$

for $L_j(z) \equiv L(k(z)'\gamma + t(z)'\lambda_j)$, and $L_j'(z) \equiv L'(k(z)'\gamma + t(z)'\lambda_j)$, $j = s, a$. In addition,

$$\begin{aligned}
G_{q,\theta_0}(x) \equiv& G^{'}(G^{-1}(F_{X|R}(x|1)))^{-1}\cdot\mathbb{E}\left[\frac{1-R}{Q_0}\cdot\nabla_{L,\theta_0}(Z)1(X\le x)\right], && (2.7.8)\\
G_{p,\theta_0}(x) \equiv& \,\mathbb{E}\left[\frac{1-R}{Q_0}\nabla_{L,\theta_0}(Z)1(X\le x)\right]\Big/ f_{X|R}(x|1) \\
&+(G(x)-F_{X|R=1}(x))\,\mathbb{E}\left[\frac{1-R}{Q_0}\nabla_{L,\theta_0}(Z)I_{b_x}K_{b_x}(X-x)\right]\Big/ f_{X|R}(x|1)^2, \\
\psi_{g,q}(a;\theta_0,G) \equiv& \mathbb{E}\left[\frac{1-R}{1-Q_0}\cdot\frac{\ell(Z)\Lambda_x(W;\beta_0)}{G'\left(G^{-1}(F_{X|R}(X|1))\right)}\right. \\
&\left.\cdot\left(\frac{(1-r)\ell(z)1(x\le X)}{1-Q_0}-\frac{rF_{X|R}(X|1)}{Q_0}\right)\right], && (2.7.9)\\
\psi_{g,p}(a;\theta_0,G) \equiv& \mathbb{E}\left[\frac{(1-R)\ell(Z)\Lambda_x(W;\beta_0)}{(1-Q_0)f_{X|R}(X|1)}\cdot\left(\frac{(1-r)\ell(z)1(x\le X)}{1-Q_0}-F_{X|R}(X|1)\right)\right] \\
&+\frac{(1-r)\ell(z)}{1-Q_0}\pi(x)-\mathbb{E}\left[\frac{(1-R)\ell(Z)}{1-Q_0}\pi(X)\right] \\
&-\frac{r-Q_0}{Q_0}\cdot\mathbb{E}\left[\frac{(1-R)\ell(Z)\Lambda_x(W;\beta_0)}{1-Q_0}\cdot\frac{G(X)}{f_{X|R}(X|1)}\right], && (2.7.10)\\
\pi(x) \equiv& \mathbb{E}\left[\Lambda_x(W;\beta_0)|X=x,R=1\right]\frac{G(x)-F_{X|R}(x|1)}{f_{X|R}(x|1)}. && (2.7.11)
\end{aligned}$$

2.8 Supplementary Appendix

This supplemental appendix contains (i) proofs of the results in Section 2.4 of the main text, (ii) auxiliary lemmas along with their proofs, and (iii) additional details of the variance estimators.

Notation: We write $\|f\|_\infty$ to denote the sup norm of f. Let $N(\varepsilon,\mathscr{F},L_r(Q))$ denote the covering number of $\mathscr{F}$ relative to the $L_r(Q)$-norm. Given an envelop function F of $\mathscr{F}$, the uniform entropy numbers and uniform entropy integral, relative to $L_r(Q)$-norm are then defined as $\sup_Q \log N(\varepsilon\|F\|_{Q,r},\mathscr{F},L_r(Q))$ and $\int_0^\delta \sup_Q \log N(\varepsilon\|F\|_{Q,r},\mathscr{F},L_r(Q))$, respectively. We say that the class $\mathscr{F}$ has bounded uniform entropy integral (BUEI) with envelop F if the uniform entropy integral is finite. K_j, $j\in\{1,2,...\}$ are finite positive constants. Let CLT, CMT, DCT, LLN, and MVT refer to the central limit theorem, the continuous mapping theorem, the dominated convergence theorem, the law of large numbers, and the mean value theorem, respectively.

2.8.1 Proofs of Lemmas and Theorems from Main Text

Proof of Lemma 2.2: The proof is divided into two steps. In the fisrt step, we show that $\widehat{\beta} \xrightarrow{p} \beta_0$.

Observe that the three-stage estimation procedure is equivalent to a GMM estimator with $(d_\gamma + 2d_\lambda + d_\beta)$ moment conditions. We collect the moment conditions by

$$m(a;\theta,q_\tau) \equiv \begin{pmatrix} m_1(a;\gamma,q_\tau) \\ m_2(a;\gamma,\lambda_s,q_\tau) \\ m_3(a;\gamma,\lambda_a,q_\tau) \\ m_4(a;\gamma,\lambda_s,\lambda_a,\beta,q_\tau) \end{pmatrix},$$

where

$$\begin{aligned}
m_1(a;\gamma,q_\tau) &\equiv \frac{r - L(k(z)'\gamma)}{L(k(z)'\gamma)(1-L(k(z)'\gamma))} L'(k(z)'\gamma)k(z), \\
m_2(a;\gamma,\lambda_s,q_\tau) &\equiv \left(\frac{r}{L(k(z)'\gamma + t(z)'\lambda_s)} - 1\right) L(k(z)'\gamma)t(z), \\
m_3(a;\gamma,\lambda_a,q_\tau) &\equiv \left(\frac{1-r}{1-L(k(z)'\gamma + t(z)'\lambda_a)} - 1\right) L(k(z)'\gamma)t(z), \\
m_4(a;\theta,q_\tau) &\equiv \left(\frac{rL(k(z)'\gamma)e(z)}{L(k(z)'\gamma + t(z)'\lambda_s)} 1(y \le q_\tau) - \frac{(1-r)L(k(z)'\gamma)e(z)}{1-L(k(z)'\gamma + t(z)'\lambda_a)} \Lambda(w;\beta)\right).
\end{aligned}$$

In addition, let

$$\begin{aligned}
\widehat{\mathscr{L}_n}(\theta) &\equiv \|\mathbb{E}_n[m(A;\theta,\widehat{q}_\tau)]\|^2_{\widetilde{\Omega}_n}, & \widetilde{\mathscr{L}_n}(\theta) &\equiv \|\mathbb{E}_n[m(A;\theta,\widehat{q}_\tau)]\|^2_{\widetilde{\Omega}}, \\
\mathscr{L}_n(\theta) &\equiv \|\mathbb{E}_n[m(A;\theta,q_\tau)]\|^2_{\widetilde{\Omega}}, & \mathscr{L}^*(\theta) &\equiv \|\mathbb{E}[m(A;\theta,q_\tau)]\|^2_{\widetilde{\Omega}},
\end{aligned}$$

where $\widetilde{\Omega}_n \equiv diag(I_{d_\gamma + 2d_\lambda}, \Omega_n)$ and $\widetilde{\Omega} \equiv diag(I_{d_\gamma + 2d_\lambda}, \Omega)$.

First, by Assumptions 2.7(c)(iii), (d), (e)(iii), (f), Lemma 2.5, and the uniform LLN,

$$\begin{aligned}
\left\|\widehat{\mathscr{L}_n}(\theta) - \widetilde{\mathscr{L}_n}(\theta)\right\| &\le \|\Omega_n - \Omega\| \cdot (\|\mathbb{E}_n[m(A;\theta,\widehat{q}_\tau)] - \mathbb{E}[m(A;\theta,q_\tau)]\|^2 + \|\mathbb{E}[m(A;\theta,q_\tau)]\|^2) \\
&= O_p(\delta_{\omega,n})(o_p(1) + O_p(1)) = o_p(1). \qquad (2.8.1)
\end{aligned}$$

Next, by the definition of $\widetilde{\mathscr{L}_n}(\theta)$ and $\mathscr{L}_n(\theta)$,

$$\left\|\widetilde{\mathscr{L}_n}(\theta) - \mathscr{L}_n(\theta)\right\|$$

$$\begin{aligned}
&\leq \|\mathbb{E}_n[m_4(A;\theta,\widehat{q}_\tau)-m_4(A;\theta,q_\tau)]\|_\Omega^2+O_p(1)\lambda_{max}(\Omega)\,\|\mathbb{E}_n[m_4(A;\theta,\widehat{q}_\tau)-m_4(A;\theta,q_\tau)]\| \\
&= o_p(1), \qquad (2.8.2)
\end{aligned}$$

where the last line follows by Assumption 2.7(d) and Lemma 2.5.

Again, by the definition of $\mathscr{L}_n(\theta)$ and $\mathscr{L}^*(\theta)$,

$$\begin{aligned}
\|\mathscr{L}_n(\theta)-\mathscr{L}^*(\theta)\| \leq &\lambda_{max}(\Omega)(\|\mathbb{E}_n[m(A;\theta,q_\tau)]-\mathbb{E}[m(A;\theta,q_\tau)]\|^2 \\
&+O_p(1)\,\|\mathbb{E}_n[m(A;\theta,q_\tau)]-\mathbb{E}[m(A;\theta,q_\tau)]\|)=o_p(1), \qquad (2.8.3)
\end{aligned}$$

where the last equality follows by Assumption 2.7(d) and Lemma 2.5.

Let $\widehat{\theta}\equiv\arg\min_{\theta\in\Theta}\widehat{\mathscr{L}_n}(\theta)$. To show the consistency of $\widehat{\theta}$, we note that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\widehat{\theta}-\theta_0\right\|\geq\varepsilon\right)\leq\mathbb{P}\left(\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon}\widehat{\mathscr{L}_n}(\theta)\leq\widehat{\mathscr{L}_n}(\theta_0)\right) \\
&\leq\mathbb{P}\left(\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon}\widetilde{\mathscr{L}_n}(\theta)\leq\widetilde{\mathscr{L}_n}(\theta_0)+2\sup_{\theta\in\Theta}|\widehat{\mathscr{L}_n}(\theta)-\widetilde{\mathscr{L}_n}(\theta)|\right) \\
&\leq\mathbb{P}\left(\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon}\mathscr{L}_n(\theta)\leq\mathscr{L}_n(\theta_0)+2\sup_{\theta\in\Theta}|\widetilde{\mathscr{L}_n}(\theta)-\mathscr{L}_n(\theta)|+o_p(1)\right) \\
&\leq\mathbb{P}\left(\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon}\mathscr{L}^*(\theta)\leq\mathscr{L}^*(\theta_0)+2\sup_{\theta\in\Theta}|\mathscr{L}_n(\theta)-\mathscr{L}^*(\theta)|+o_p(1)\right) \\
&\leq\mathbb{P}\left(\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon}\|\mathbb{E}[m(A;\theta,q_\tau)]\|^2\leq\frac{o_p(1)}{\lambda_{min}(\Omega)}\right) \\
&=o(1),
\end{aligned}$$

where the first inequality is obtained by the definition of $\widehat{\theta}$, and the third to fifth line follow by (2.8.1)–(2.8.3), respectively. By Assumptions 2.3, 2.7(a)(ii), (c)(i), and (e)(ii), $\theta=\theta_0$ is the unique solution to $\|\mathbb{E}[m(A;\theta,q_\tau)]\|=0$. Using this fact, the last line is obtained by Exercise 5.27 in Van der Vaart (1998).

In the second step, we prove that $\sqrt{n}(\widehat{\beta}-\beta_0)\xrightarrow{d}N(0,\Sigma_\beta)$.

A first-order Taylor expansion yields that

$$\begin{aligned}
o_p(1)&=M_n'\widetilde{\Omega}_n\frac{1}{\sqrt{n}}\sum_{i=1}^n m(A_i;\widehat{\theta},\widehat{q}_\tau) \\
&=M_n'\widetilde{\Omega}_n\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n(m(A_i;\theta_0,\widehat{q}_\tau)-m(A_i;\theta_0,q_\tau)+m(A_i;\theta_0,q_\tau))+\widetilde{M}_n\sqrt{n}(\widehat{\theta}-\theta_0)\right),
\end{aligned}$$

where $M_n\equiv\mathbb{E}_n[\nabla_\theta m(A;\widehat{q}_\tau,\widehat{\theta})]$, and $\widetilde{M}_n\equiv\mathbb{E}_n[\nabla_\theta m(A;\widehat{q}_\tau,\widetilde{\theta})]$, for some $\widetilde{\theta}$ lying between $\widehat{\theta}$ and θ_0. Using $\widehat{\theta}\xrightarrow{p}\theta_0$, it can be shown that both M_n and $\widetilde{M}_n$ converge in probability to $M \equiv \mathbb{E}[\nabla_\theta m(A;\theta_0,q_\tau)]$. The proof is similar to that of Lemma 2.5, we omit the details to avoid repetition. Thus,

$$\begin{aligned}
&\sqrt{n}(\widehat{\theta}-\theta_0)\\
&= -M_\Omega^{-1}M'\widetilde{\Omega}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(A_i;\theta_0,\widehat{q}_\tau)-m(A_i;\theta_0,q_\tau)+m(A_i;\theta_0,q_\tau)\right)+o_p(1)\\
&= -M_\Omega^{-1}M'\widetilde{\Omega}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{E}[m(A;\theta_0,\widehat{q}_\tau)-m(A;\theta_0,q_\tau)|\mathfrak{Y}_n]+m(A_i;\theta_0,q_\tau)\right)+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\theta(A_i;\theta_0,q_\tau)+o_p(1), \qquad (2.8.4)
\end{aligned}$$

where

$$M_\Omega \equiv M'\widetilde{\Omega}M, \qquad (2.8.5)$$

$$\psi_\theta(a;\theta_0,q_\tau) \equiv -M_\Omega^{-1}M'\widetilde{\Omega}\left(m(a;\theta_0,q_\tau)-M_{q_\tau}\frac{r\cdot(1(y\le q_\tau)-\tau)}{Q_0 f_{Y|R}(q_\tau|1)}\right), \qquad (2.8.6)$$

and $M_{q_\tau}=(0',0',0',\mathbb{E}[e(Z)Rf_{Y|ZR}(q_\tau|Z,1)]')'$. The second equality follows from Lemma 2.4. The third one is due to a first-order Taylor expansion, the uniform LLN, and the following fact,

$$\widehat{q}_\tau - q_\tau = -\mathbb{E}_n\left[\frac{R(1(Y\le q_\tau)-\tau)}{Q_0 f_{Y|R}(q_\tau|1)}\right]+o_p(n^{-1/2}). \qquad (2.8.7)$$

See, e.g. Firpo (2007).

The asymptotic linear representation of $\widehat{\beta}$ corresponds to the last d_β-elements of (2.8.6). Let $M_{k,s}$ denote the Jacobian matrix of the k-th subvector of $\mathbb{E}[m(A;q_\tau,\theta)]$ with respect to the s-th subvector of θ evaluated at θ_0. It is straightforward to show that,

$$\sqrt{n}(\widehat{\beta}-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\beta(A_i;\theta_0,q_\tau),$$

where

$$\begin{aligned}
\psi_\beta(a;\theta_0,q_\tau) \equiv -\frac{1}{\sqrt{n}}(M'_{44}\Omega M_{44})^{-1}&\left\{M'_{44}\Omega M_{4,-4}S^{-1}M'_{-4,-4}\begin{pmatrix} m_1(a;\gamma_0)\\ m_2(a;\gamma_0,\lambda_{s,0})\\ m_3(a;\gamma_0,\lambda_{a,0})\end{pmatrix}\right.\\
&\left.-M'_{44}\Omega^{1/2}(I-\Omega^{1/2}M_{4,-4}S^{-1}M'_{4,-4}\Omega^{1/2}\mathscr{M}_{\Omega^{1/2}M_{4,4}})\Omega^{1/2}\widetilde{m}_4(a;\theta_0,q_\tau)\right\}, \qquad (2.8.8)
\end{aligned}$$

for $M_{4,-4} \equiv (M_{41}, M_{42}, M_{43})$, $S \equiv M'_{-4,-4}M_{-4,-4} - M'_{4,-4}\Omega^{1/2}\mathscr{M}_{\Omega^{1/2}M_{4,4}}\Omega^{1/2}M_{4,-4}$, $\mathscr{M}_H \equiv I - H(H'H)^{-1}H'$, and

$$\widetilde{m}_4(a;\theta_0,q_\tau) \equiv m_4(a;\theta_0,q_\tau) - \mathbb{E}[Rf_{Y|ZR}(q_\tau|Z,1)e(Z)] \cdot \frac{r\cdot(1(y \leq q_\tau) - \tau)}{Q_0 f_{Y|R}(q_\tau|1)}.$$

This concludes the proof of Lemma 2.2. ■

Proof of Theorem 2.3: Part I: Asymptotic result for $\widehat{UQE}_q$.

Decomposing the difference, we have

$$\begin{aligned}
\widehat{UQE}_q(\tau,G) - UQE_q(\tau,G) &\leq -\left(\frac{1}{\widehat{f}_{Y|R}(\widehat{q}_\tau|1)} - \frac{1}{f_{Y|R}(q_\tau|1)}\right)\widehat{d}_{q,n}(\widehat{\theta},G) \\
&- \frac{1}{f_{Y|R}(q_\tau|1)}(\widehat{d}_{q,n}(\widehat{\theta},G) - \widehat{d}_{q,n}(\theta_0,G)) \\
&- \frac{1}{f_{Y|R}(q_\tau|1)}\widehat{d}_{q,n}(\theta_0,G) - UQE(\tau,G) \\
&\equiv \Delta_{q,1} + \Delta_{q,2} + \Delta_{q,3},
\end{aligned} \tag{2.8.9}$$

where

$$\begin{aligned}
\widehat{d}_{q,n}(\widehat{\theta},G) &\equiv \mathbb{E}_{n_a}[\widehat{\ell}(Z)\Lambda_x(W;\widehat{\beta})\widehat{g}_q(X)], \qquad (2.8.10)\\
\widehat{d}_{q,n}(\theta_0,G) &\equiv \frac{1}{1-Q_0}\mathbb{E}_n[(1-R)\ell(Z)\Lambda_x(W;\beta_0)\widehat{g}_q(X;\theta_0)],
\end{aligned}$$

and $\widehat{g}_q(x;\theta_0) \equiv G^{-1}\left(\frac{1}{1-Q_0}\mathbb{E}_n[(1-R)\ell(Z)1(X \leq x)]\right) - x$.

We proceed by deriving the asymptotic linear representation of each term.

For $\Delta_{q,1}$, we focus on the inverse of the density estimate first,

$$\frac{1}{\widehat{f}_{Y|R}(\widehat{q}_\tau|1)} - \frac{1}{f_{Y|R}(q_\tau|1)} = -\frac{\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - f_{Y|R}(q_\tau|1)}{f_{Y|R}(q_\tau|1)^2} + \xi_1, \tag{2.8.11}$$

where

$$\xi_1 \equiv \frac{(\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - f_{Y|R}(q_\tau|1))^2}{f_{Y|R}(q_\tau|1)^2\widehat{f}_{Y|R}(\widehat{q}_\tau|1)}.$$

Next, we establish the rate for $\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - f_{Y|R}(q_\tau|1)$. Decomposing the difference,

$$\begin{aligned}\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - f_{Y|R}(q_\tau|1) =& (\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - \widetilde{f}_{Y|R}(\widehat{q}_\tau|1)) + (\widetilde{f}_{Y|R}(\widehat{q}_\tau|1) - \mathbb{E}[\widetilde{f}_{Y|R}(\widehat{q}_\tau|1)]) \\ &+ (\mathbb{E}[\widetilde{f}_{Y|R}(\widehat{q}_\tau|1)] - f_{Y|R}(\widehat{q}_\tau|1)) + (f_{Y|R}(\widehat{q}_\tau|1) - f_{Y|R}(q_\tau|1)) \\ \equiv& \Delta_{f,1} + \Delta_{f,2} + \Delta_{f,3} + \Delta_{f,4},\end{aligned}$$

where $\widetilde{f}_{Y|R}(q_\tau|1) \equiv \mathbb{E}_n\left[RK_{b_y}(Y - q_\tau)\right]/Q_0$.

We analyze each term in turn. For $\Delta_{f,1}$, a first-order approximation and applying the uniform LLN yield,

$$\begin{aligned}\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - \widetilde{f}_{Y|R}(\widehat{q}_\tau|1) &= -\frac{\mathbb{E}_n[R - Q_0]}{Q_0} \cdot \frac{\mathbb{E}_n[RK_{b_y}(Y - \widehat{q}_\tau)]}{\mathbb{E}_n[R]} \\ &= -\frac{\mathbb{E}_n[R - Q_0]}{Q_0} \cdot \left(\frac{\mathbb{E}[RK_{b_y}(Y - q_\tau)]}{Q_0} + \frac{\mathbb{E}_n[RK_{b_y}(Y - \widehat{q}_\tau)]}{\mathbb{E}_n[R]} - \frac{\mathbb{E}[RK_{b_y}(Y - q_\tau)]}{Q_0}\right) \\ &= -\frac{\mathbb{E}_n[R - Q_0]}{Q_0} \cdot \frac{\mathbb{E}[RK_{b_y}(Y - q_\tau)]}{Q_0} + o_p(n^{-1/2}),\end{aligned}$$

where the third line follows along a similar line of argument as in Section B.3.2 of Sasaki et al. (2022).

Likewise, we have that

$$\Delta_{f,2} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{Q_0}\left\{R_i K_{b_y}(Y_i - q_\tau) - \mathbb{E}[RK_{b_y}(Y - q_\tau)]\right\} + o_p(n^{-1/2}).$$

Lastly, we bound the bias,

$$\begin{aligned}\Delta_{f,3} &= \mathbb{E}\left[\frac{RK_{b_y}(Y - \widehat{q}_\tau)}{Q_0}\right] - f_{Y|R}(\widehat{q}_\tau|1) \\ &= \frac{b_y^2(f''_{Y|R}(q_\tau|1) + (f''_{Y|R}(\widetilde{q}_\tau|1) - f''_{Y|R}(q_\tau|1)))}{2}\int y^2 K_y(y)dy \\ &= B_{s,y}(q_\tau, b_y) + o_p(b_y^2), \qquad (2.8.12)\end{aligned}$$

where $B_{s,y}(q_\tau, b_y) \equiv 0.5 b_y^2 f''_{Y|R}(q_\tau|1)\int y^2 K_y(y)dy$ and $\widetilde{q}$ lies between $\widehat{q}_\tau$ and $\widehat{q}_\tau + c_y b_y$, for some $|c_y| < \bar{y}$, where $\bar{y}$ is the maximum absolute value of $\text{Supp}(K_y)$. The second line follows by changing variables and a second-order expansion of $f_{Y|R}(\cdot|1)$ about $\widehat{q}_\tau$. The rate of remainder is $o_p(b_y^2)$ because $|f''_{Y|R}(\widetilde{q}_\tau|1) - f''_{Y|R}(q_\tau|1)| \le |f''_{Y|R}(\widetilde{q}_\tau|1) - f''_{Y|R}(\widehat{q}_\tau|1)| + |f''_{Y|R}(\widehat{q}_\tau|1) - f''_{Y|R}(q_\tau|1)| = O_p(b_y + n^{-1/2})$, and by Assumption 2.8(c).

In view of (2.8.7), it follows that

$$\Delta_{f,4} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i f'_{Y|R}(q_\tau|1)(1(Y_i \leq q_\tau) - \tau)}{Q_0 f_{Y|R}(q_\tau|1)} + o_p(n^{-1/2}).$$

Collecting the linear expansions of $\Delta_{f,1}$–$\Delta_{f,4}$, we get

$$\begin{aligned}
&\widehat{f}_{Y|R}(\widehat{q}_\tau|1) - f_{Y|R}(q_\tau|1) \\
=&\mathbb{E}_n\left[\frac{R}{Q_0}\left\{K_{b_y}(Y-q_\tau) - \frac{\mathbb{E}[RK_{b_y}(Y-q_\tau)]}{Q_0} - \frac{(1(Y \leq q_\tau) - \tau)f'_{Y|R}(q_\tau|1)}{f_{Y|R}(q_\tau|1)}\right\}\right] \\
&+ B_{s,y}(q_\tau, b_y) + \xi_2,
\end{aligned} \tag{2.8.13}$$

where ξ_2 collects the remainder terms from each $\Delta_{f,j}$, for $j = 1, ..., 4$, and $\xi_2 = o_p(n^{-1/2}b_y^{-1/2} + b_y^2)$.

By Lemma 2.6, we have

$$\widehat{d}_{q,n}(\widehat{\theta}, G) \xrightarrow{p} d_q(\theta_0, G). \tag{2.8.14}$$

Collecting the results in (2.8.11), (2.8.13), and (2.8.14), we deduce that

$$\Delta_{q,1} = \frac{1}{n}\sum_{i=1}^{n} \psi_{q_1}(A_i, \theta_0, q_\tau, b_y) + \frac{B_{s,y}(q_\tau, b_y) d_q(\theta_0, G)}{f^2_{Y|R}(q_\tau|1)} + o_p(n_s^{-1/2}b_y^{-1/2} + b_y^2),$$

where

$$\begin{aligned}
\psi_{q_1}(a, \theta_0, q_\tau, b_y) \equiv \Bigg\{ \frac{r}{Q_0}\Big(&K_{b_y}(y-q_\tau) - \mathbb{E}[K_{b_y}(Y-q_\tau)|R=1] \\
&- \frac{(1(y \leq q_\tau) - \tau)f'_{Y|R}(q_\tau|1)}{f_{Y|R}(q_\tau|1)}\Big) \cdot d_q(\theta_0, G)\Bigg\} \Big/ f^2_{Y|R}(q_\tau|1).
\end{aligned} \tag{2.8.15}$$

Next, we derive the asymptotically linear representation of $\Delta_{q,2}$. Decomposing the difference, we get

$$\begin{aligned}
\widehat{d}_{q,n}(\widehat{\theta}, G) - \widehat{d}_{q,n}(\theta_0, G) &= \Big(\widehat{d}_{q,n}(\widehat{\theta}, G) - \widetilde{d}_{q,n}(\widehat{\theta}, G)\Big) + \Big(\widetilde{d}_{q,n}(\widehat{\theta}, G) - \widehat{d}_{q,n}(\theta_0, G)\Big) \\
&\equiv \Delta_{q,4} + \Delta_{q,5},
\end{aligned}$$

where $\widetilde{d}_{q,n}(\widehat{\theta}, G) \equiv \mathbb{E}_n\left[\frac{(1-R)L(k(Z)'\widehat{\gamma} + t(Z)'\widehat{\lambda}_s)}{Q_0(1-L(k(Z)'\widehat{\gamma} + t(Z)'\widehat{\lambda}_a))}\Lambda_x(W;\widehat{\beta})\widehat{g}_q(X)\right]$. By a second-order Taylor expansion with respect to $\mathbb{E}_n[R]$ around Q_0,

$$\Delta_{q,4} = -\frac{\mathbb{E}_n[R-Q_0]}{Q_0} \cdot (d_q(\theta_0, G) + \widetilde{d}_{q,n}(\widehat{\theta}, G) - d_q(\theta_0, G)) + O_p(n^{-1})$$

$$= -\frac{\mathbb{E}_n[R-Q_0]}{Q_0} \cdot d_q(\theta_0, G) + o_p(n^{-1/2}),$$

where the second line follows by Lemma 2.2, Assumptions 2.7(c)(ii), (e)(ii), Assumption 2.8(d), and the uniform LLN.

Again, by expanding $\widetilde{d}_{q,n}(\widehat{\theta}, G)$ around θ_0, we have that

$$\Delta_{q,5} = M_{\theta,q,n}(\widetilde{\theta})'(\widehat{\theta} - \theta_0) + o_p\left(\left\|\widehat{\theta} - \theta_0\right\|\right),$$

where $\widetilde{\theta}$ is some value between $\widehat{\theta}$ and θ_0,

$$M_{\theta,q,n}(\theta) \equiv \mathbb{E}_n\left[\frac{1-R}{Q_0}\begin{pmatrix}\Lambda_x(W;\beta)\left(\nabla_{L,\theta}(Z)\cdot\widehat{g}_q(X;\theta) + G_{q,\theta}(X)\right)\\ \frac{L_s(Z)}{1-L_a(Z)}\Lambda_{x,\beta}(W;\beta)\widehat{g}_q(X;\theta)\end{pmatrix}\right],$$

and

$$G_{q,\theta}(x) \equiv G^{'}(G^{-1}(\widehat{F}_{X|R}(x|1)))^{-1}\cdot\mathbb{E}_{n_a}\left[\nabla_{L,\theta}(Z)1(X\leq x)\right].$$

Slightly modifying Theorem 8.2 in Newey and McFadden (1994) and applying it to $M_{\theta,q,n}$, we can deduce that $M_{\theta,q,n}(\widetilde{\theta}) \xrightarrow{p} M_{\theta,q}(\theta_0)$, where $M_{\theta,q}(\theta_0)$ is defined in (2.7.6). Therefore, $\Delta_{q,5} = M_{\theta,q}(\theta_0)'(\widehat{\theta}-\theta_0) + o_p(n^{-1/2})$.

In view of (2.8.4) and (2.8.6), $\Delta_{q,5} = \frac{1}{n}\sum_{i=1}^n M_{\theta,q}(\theta_0)'\psi_\theta(A_i;\theta_0,q_\tau) + o_p(n^{-1/2})$. Hence, $\Delta_{q,2} = \frac{1}{n}\sum_{i=1}^n \psi_{q_2}(A_i;\theta_0, q_\tau) + o_p(n^{-1/2})$, where

$$\psi_{q_2}(a;\theta_0,q_\tau) \equiv -\frac{1}{f_{Y|R}(q_\tau|1)}\left(M_{\theta,q}(\theta_0)'\psi_\theta(a;\theta_0,q_\tau) - \frac{r-Q_0}{Q_0}d_q(\theta_0,G)\right). \tag{2.8.16}$$

Next, we derive the asymptotic linear expansion of $\Delta_{q,3}$. By definition,

$$\begin{aligned}\widehat{d}_{q,n}(\theta_0,G) - d_q(\theta_0,G) =& \mathbb{E}_n[\Delta\phi_q(A;\theta_0,G)]\\ &+\left(\frac{1}{1-Q_0}\mathbb{E}_n[(1-R)\ell(Z)\Lambda_x(W;\beta_0)g_q(X)] - d_q(\theta_0,G)\right),\end{aligned}$$

where

$$\Delta\phi_q(a;\theta_0,G) \equiv \frac{1-r}{1-Q_0}\ell(z)\Lambda_x(w;\beta_0)(\widehat{g}_q(x;\theta_0) - g_q(x;\theta_0)). \tag{2.8.17}$$

By Lemma 2.7, we deduce that $\mathbb{E}_n[\Delta\phi_q(A;\theta_0,G)] = \frac{1}{n}\sum_{i=1}^n \psi_{g,q}(A_i;\theta_0,G) + o_p(n^{-1/2})$, where $\psi_{g,q}$ is given in (2.7.9).

Using this result, by the definition of $\widehat{d}_{q,n}(\theta_0, G)$, and by Theorem 2.1, it follows that $\Delta_{q,3} = \frac{1}{n}\sum_{i=1}^{n} \psi_{q_3}(A_i; \theta_0, q_\tau)$, where

$$\psi_{q_3}(a; \theta_0, q_\tau) \equiv -\frac{1}{f_{Y|R}(q_\tau|1)}\left(\psi_{g,q}(a; \theta_0, G) + \frac{1-r}{1-Q_0}\ell(z)\Lambda_x(w; \beta_0)g_q(x)\right.$$
$$\left. - d_q(\theta_0, G)\right). \quad (2.8.18)$$

Collecting the linear expansions associated with $\Delta_{q,j}$, $j = 1, 2, 3$, and by the Lyapunov CLT, we deduce that, with ψ_q given in (2.7.3),

$$\sqrt{nb_y}(\widehat{UQE}_q(\tau, G) - UQE_q(\tau, G)) \xrightarrow{d} N(0, \lim_{b_y \to 0} b_y \mathbb{E}[\psi_q(A; \theta_0, q_\tau, b_y)\psi_q(A; \theta_0, q_\tau, b_y)']).$$

To finish the proof, we need to verify the asymptotic variance of $\widehat{UQE}_q(\tau, G)$,

$$\begin{aligned}
&\lim_{b_y \to 0} b_y \mathbb{E}[\psi_q(A; \theta_0, q_\tau, b_y)\psi_q(A; \theta_0, q_\tau, b_y)'] \\
&= \frac{D^2(\theta_0, G)}{f^4_{Y|R}(q_\tau|1)} \cdot \lim_{b_y \to 0}\left\{\mathbb{E}\left[\frac{b_y}{Q_0}K^2_{b_y}(Y - q_\tau)|R = 1\right] - \frac{b_y}{Q_0}\left(\mathbb{E}[K_{b_y}(Y - q_\tau)|R = 1]\right)^2\right\} \\
&= \frac{D^2(\theta_0, G)}{f^4_{Y|R}(q_\tau|1)Q_0} \cdot \lim_{b_y \to 0}\left\{\int \frac{1}{b_y}K^2_y\left(\frac{y - q_\tau}{b_y}\right)f_{Y|R}(y|1)dy - b_y\left(f_{Y|R}(q_\tau) + o(b_y^2)\right)^2\right\} \\
&= \frac{D^2(\theta_0, G)}{f^4_{Y|R}(q_\tau|1)Q_0} \cdot \lim_{b_y \to 0}\left\{\int K^2_y(u) f_{Y|R}(q_\tau|1)du + O(b_y)\right\} \\
&= \frac{D^2(\theta_0, G)}{f^3_{Y|R}(q_\tau|1)Q_0} \cdot \int K^2_y(u)\, du,
\end{aligned}$$

where the third line is due to (2.8.12), and the fourth one is obtained by changing variable, a first-order expansion of K_y, and by Assumption 2.8(b).

Part II: Asymptotic results for $\widehat{UQE}_p$.

Let $\widehat{d}_{p,n}(\theta_0, G) \equiv \frac{1}{1-Q_0}\mathbb{E}_n[(1-R)\ell(Z)\Lambda_x(W; \beta_0)\widehat{g}_p(X; \theta_0)]$, we perform a decomposition analogous to (2.8.9),

$$\begin{aligned}
\widehat{UQE}_p(\tau, G) - UQE_p(\tau, G) \leq &-\left(\frac{1}{\widehat{f}_{Y|R}(\widehat{q}_\tau|1)} - \frac{1}{f_{Y|R}(q_\tau|1)}\right)\widehat{d}_{p,n}(\widehat{\theta}, G) \\
&- \frac{1}{f_{Y|R}(q_\tau|1)}(\widehat{d}_{p,n}(\widehat{\theta}, G) - \widehat{d}_{p,n}(\theta_0, G)) \\
&- \frac{1}{f_{Y|R}(q_\tau|1)}\widehat{d}_{p,n}(\theta_0, G) - UQE(\tau, G) \\
\equiv &\Delta_{p,1} + \Delta_{p,2} + \Delta_{p,3},
\end{aligned} \quad (2.8.19)$$

with influence functions associated with $\Delta_{p,j}$ denoted by ψ_{pj}, $j=1,2,3$, respectively. The first term represents the estimation error of $f_{Y_s|R=1}$, the second term corresponds to the estimation effect of θ, and the last term accounts for the contribution of $\widehat{g}_p$.

The first term can be treated in a similar fashion as in the previous part, and therefore, we give the linear expansion directly without a proof,

$$\begin{aligned}\psi_{p_1}(a;\theta_0,q_\tau) \equiv \Bigg\{ & \frac{r}{Q_0}\Big(K_{b_y}(y-q_\tau) - \mathbb{E}[K_{b_y}(Y-q_\tau)|R=1] \\ & - \frac{(1(y\le q_\tau)-\tau)f'_{Y|R}(q_\tau|1)}{f_{Y|R}(q_\tau|1)}\Big)\cdot d_p(\theta_0,G)\Bigg\} \Bigg/ f^2_{Y|R}(q_\tau|1). \end{aligned} \tag{2.8.20}$$

The linear expansion of $\Delta_{p,2}$ also takes a similar form as in (2.8.16), with

$$\psi_{p_2}(a;\theta_0,q_\tau) \equiv -\frac{1}{f_{Y|R}(q_\tau|1)}\left(M_{\theta,p}(\theta_0)'\psi_\theta(a;\theta_0,q_\tau) - \frac{r-Q_0}{Q_0}\cdot d_p(\theta_0,G)\right), \tag{2.8.21}$$

where $M_{\theta,p}(\theta_0)$ is defined in (2.7.6). Assumption 2.9(c) allows us to ignore the bias in the approximation of $G_{p,\theta_0}(x)$ and in turn, $M_{\theta,p}(\theta_0)$.

Next, we apply Lemma 5.1 in Newey (1994) to derive the linear expansion of $\Delta_{p,3}$. Define $\rho(\cdot)=(\rho_1(\cdot),\rho_2(\cdot),\rho_3)$,

$$\begin{aligned}\phi_p(a;\rho) &\equiv -\frac{(1-r)\ell(z)\Lambda_x(w;\beta_0)}{1-Q_0}\cdot\frac{(G(x)-\rho_1(x)/\rho_3)}{\rho_2(x)/\rho_3} - d_p(\theta_0,G),\\ \Phi_{p,1}(a;\rho_1) &\equiv \frac{(1-r)\ell(z)\Lambda_x(w;\beta_0)}{1-Q_0}\cdot\frac{\rho_1(x)}{f_{X|R}(x|1)},\\ \Phi_{p,2}(a;\rho_2) &\equiv \frac{(1-r)\ell(z)\Lambda_x(w;\beta_0)}{1-Q_0}\cdot\frac{(G(x)-F_{X|R}(x|1))\rho_2(x)}{f^2_{X|R}(x|1)},\\ \Phi_{p,3}(a;\rho_3) &\equiv -\frac{(1-r)\ell(z)\Lambda_x(w;\beta_0)}{1-Q_0}\cdot\frac{G(x)\rho_3}{f_{X|R}(x|1)},\end{aligned}$$

and $\Phi_p(a;\rho)\equiv\sum_{j=1}^3\Phi_{p,j}(a;\rho_j)$. With these notations in hand, we can rewrite $\Delta_{p,3}$ as follows,

$$\Delta_{p,3} = -\frac{1}{f_{Y|R}(q_\tau|1)}(\mathbb{E}_n[\phi_p(A;\widehat{\rho}_n)] - \mathbb{E}[\phi_p(A;\rho_0)]),$$

where $\widehat{\rho}_n(\cdot)\equiv\left(\mathbb{E}_n\left[\frac{(1-R)\ell(Z)1(X\le\cdot)}{1-Q_0}\right],\mathbb{E}_n\left[\frac{(1-R)\ell(Z)I_{b_x}K_{b_x}(X-\cdot)}{1-Q_0}\right],\frac{E_n[R]}{Q_0}\right)'$, and $\rho_0(\cdot)\equiv(F_{X|R}(\cdot|1),f_{X|R}(\cdot|1),1)'$. We proceed to verify the conditions of Lemma 5.1 in Newey (1994).

The first set of conditions controls the linearization error. By directly bounding the difference, we can show that

$$\left\|\phi_p(a;\rho)-\phi_p(a;\rho_0)-\Phi_p(a;\rho-\rho_0)\right\| \leq \sum_{j=1,2} K_j \sup_{x\in\mathscr{X}} |\rho_j(x)-\rho_{j,0}(x)|^2 + K_3|\rho_3-\rho_{3,0}|^2, \tag{2.8.22}$$

where the inequality holds under Assumption 2.1(d)(i), Assumptions 2.7(c), (e), Assumption 2.8(d), and Assumption 2.9(a). By (2.8.22), Assumption 5.1(ii) in Newey (1994) is satisfied if $\sqrt{n}\,\|\widehat{\rho}-\rho_0\|_\infty \xrightarrow{p} 0$. In what follows, we provide the proof of this convergence result.

First, by Theorem B in Section 2.1.4 in Serfling (2009), we have $\sqrt{n}\,\|\widehat{\rho}_1-\rho_{1,0}\|_\infty^2 = O_p(n^{-1/2}\log(\log(n))) = o_p(1)$.

Next, we let

$$\widetilde{\rho}_2(x) \equiv \mathbb{E}\left[\frac{(1-R)\ell(Z)I_{b_x}K_{b_x}(X-x)}{1-Q_0}\right].$$

By the triangular inequality, $\|\widehat{\rho}_2-\rho_{2,0}\|_\infty \leq \|\widehat{\rho}_2-\widetilde{\rho}_2\|_\infty + \|\widetilde{\rho}_2-\rho_{2,0}\|_\infty$. Under the rate condition in Assumption 2.9(c), Lemma 8.10 in Newey and McFadden (1994) yields that $\|\widehat{\rho}_2-\widetilde{\rho}_2\|_\infty = O_p(\log(n)^{1/2}n^{-1/2}b_x^{-1/2})$.

Next, we bound the bias, $\widetilde{\rho}_2-\rho_{2,0}$,

$$\begin{aligned}
\widetilde{\rho}_2(x) =& \int_{\underline{x}}^{\bar{x}} I_{b_x}K_{b_x}(v-x)f_{X|R}(v|1)dv \\
=& \int_{I_x} K_x(u)f_{X|R}(x+ub_x|1)du \\
=& f_{X|R}(x|1)\int_{I_x} K_x(u)du + b_x f'_{X|R}(x|1)\int_{I_x} uK_x(u)du \\
&+ \frac{b_x^2 f''_{X|R}(\widetilde{x}|1)}{2}\int_{I_x} u^2K_x(u)du \\
=& f_{X|R}(x|1) + \frac{b_x^2 f''_{X|R}(x|1)}{2}\int_{-\infty}^{\infty} u^2K_x(u)du + o_p(b_x^2).
\end{aligned}$$

where $I_x \equiv [(\underline{x}-x)/b_x+\rho_x/2, (\bar{x}-x)/b_x-\rho_x/2]$ and $\widetilde{x}$ is some value between x and $x+c_xb_x$, with $|c_x| \leq \max\{|\underline{x}|,|\bar{x}|\}$. The first line is due to Assumptions 2.1(c) and (e), the second line follows by changing variable, and the last line is due to Assumptions 2.9(a) and (b). Using this fact, we deduce that $\sqrt{n}\,\|\widehat{\rho}_2-\rho_{2,0}\|_\infty^2 = O_p(\log(n)n^{-1/2}b_x^{-1}+\sqrt{n}b_x^4)$, which is $o_p(1)$ under Assumption 2.9(c).

Lastly, $\sqrt{n}\,\|\widehat{\rho}_3-\rho_{3,0}\|^2 = O_p(n^{-1/2})$. In view of the above three results, the desired condition is verified.

In the next step, we verify Assumption 5.2. Using Lemma 8.4 in Newey and McFadden (1994), the condition is satisfied so long as $\mathbb{E}[\|\Phi(a;\widehat{\rho}-\rho_0)\|^2] \xrightarrow{p} 0$. The latter condition follows by $\|\Phi(a;\rho)\| \leq K_4\,\|\rho\|_\infty$ and DCT. The constant K_4 is finite under Assumption 2.1(d)(i), Assumptions 2.7(c), (e), Assumption 2.8(d), and Assumption 2.9(a).

Finally, we need to show the mean-square continuity (Assumption 5.3 in Newey (1994)). Towards this end, we derive the asymptotic linear representation of $\int \Phi_p(a, \widehat{\rho} - \rho) dF_A(a)$. We focus on ρ_2 first. Let

$$\psi_{p,x,2}(a; \theta_0, G) \equiv \frac{(1-r)\ell(z)}{1-Q_0}\pi(x) - \mathbb{E}\left[\frac{(1-R)\ell(Z)}{1-Q_0}\pi(X)\right], \tag{2.8.23}$$

where $\pi(x)$ is defined in (2.7.11). Now we proceed to verify that $\psi_{p,x,2}$ is indeed the influence function of $\int \Phi_p(a; \rho_2 - \rho_{2,0}) dF_A(a)$. It amounts to show that

$$\left\| \int \Phi_{p,2}(a; \widehat{\rho}_2 - \rho_{2,0}) dF_A(a) - \frac{1}{n}\sum_{i=1}^{n} \psi_{p,x,2}(A_i; \theta_0, G) \right\| = o_p(n^{-1/2}).$$

For this purpose, we bound the first two moments and apply Chebyshev's inequality. For the first moment, we have

$$\begin{aligned}
&\left\| \sqrt{n}\mathbb{E}\left[\int \Phi_{p,2}(a; \widehat{\rho}_2 - \rho_{2,0}) dF_A(a) - \frac{1}{n}\sum_{i=1}^{n} \psi_{p,x,2}(A_i; \theta_0, G) \right] \right\| \\
=&\sqrt{n}\left\| \mathbb{E}\left[\int \Phi_{p,2}(a; \widehat{\rho}_2 - \rho_{2,0}) dF_A(a) \right] \right\| \\
=&\sqrt{n}\left\| \int\int_{I_{b_x}} \pi(x) f_{X|R}(v|1) K_{b_x}(v-x) dx dv - \int\int \pi(x) f_{X|R}(x|1) K_x(u) du dx \right\| \\
\leq&\sqrt{n}\left\| \int\int_{I_x} \pi(x + b_x u) f_{X|R}(x|1) K_x(u) du dx - \int\int_{I_x} \pi(x) f_{X|R}(x|1) K_x(u) du dx \right\| \\
&+ \sqrt{n}\left\| \int\int_{I_x^c} \pi(x) f_{X|R}(x|1) K_x(u) du dx \right\| \\
\leq&\sqrt{n}\left\| \int f_{X|R}(x|1) \int_{I_x^c} (\pi'(x) b_x u + \pi''(\widetilde{x}) b_x^2 u^2/2) K_x(u) du dx \right\| \\
&+ \left\| \pi(x) f_{X|R}(x|1) \right\|_\infty \cdot \sqrt{n} \left\| \int_{I_x^c} K_x(u) du \right\| \\
=&O(n^{1/2} b_x^2) + o(1) = o(1),
\end{aligned} \tag{2.8.24}$$

where I_x^c is the complement of I_x relative to $\mathscr{X}$, and $\widetilde{x}$ is some value between x and $x + c_x b_x$. The second equality follows because

$$\begin{aligned}
&\int \Phi_{p,2}(a; \rho_2) dF_A(a) \\
&\qquad = \int\int \ell(z) \Lambda_x(w; \beta_0) \frac{(G(x) - F_{X|R}(x|1))\rho_2(x)}{f_{X|R}^2(x|1)} f_{X|ZR}(x|z,0) f_{Z|R}(z|0) dz dx \\
&\qquad = \int\int \Lambda_x(w; \beta_0) \frac{(G(x) - F_{X|R}(x|1))\rho_2(x)}{f_{X|R}^2(x|1)} f_{X|ZR}(x|z,1) f_{Z|R}(z|1) dz dx
\end{aligned}$$

$$
\begin{aligned}
&= \int \left(\int \Lambda_x(w;\beta_0) f_{Z|XR}(z|x,1)dz \right) \frac{(G(x)-F_{X|R}(x|1))\rho_2(x)}{f_{X|R}(x|1)}dx \\
&= \int \mathbb{E}[\Lambda_x(W;\beta_0)|X=x,R=1] \frac{(G(x)-F_{X|R}(x|1))\rho_2(x)}{f_{X|R}(x|1)}dx \\
&= \int \pi(x)\rho_2(x)dx.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\int \Phi_{p,2}(a;\widehat{\rho}_2)dF_A(a)\right] &= \mathbb{E}\left[\frac{(1-R)\ell(Z)}{1-Q}\int \pi(x) I_{b_x} K_{b_x}(X-x)dx\right] \\
&= \int\int \pi(x) I_{b_x} K_{b_x}(v-x) f_{X|R}(v|1)dxdv, \\
\mathbb{E}\left[\int \Phi_{p,2}(a;\rho_{2,0})dF_A(a)\right] &= \int \pi(x) f_{X|R}(x|1)dx \cdot \int K_x(u)du \\
&= \int\int \pi(x) K_x(u) f_{X|R}(x|1)dudx.
\end{aligned}
$$

The second inequality of (2.8.24) follows from a second-order Taylor expansion with respect to $\pi(\cdot)$, which is valid under Assumption 2.7(e) and Assumption 2.9(a). The second-to-last equality of (2.8.24) is due to Assumptions 2.9(a) and (b).

Next, since $\pi(\cdot)$ is continuous and bounded, and K_x has a compact support, we have that, by DCT, $\int_{I_x} \pi(x - b_x u)K_x(u)du \to \pi(x)$. As a consequence, for the second moment,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sqrt{n}\int \Phi_{p,2}(a;\widehat{\rho}_2-\rho_{2,0})dF_A(a) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{p,x,2}(A_i;\theta_0,G)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\frac{(1-R)\ell(Z)}{1-Q_0}\left(\int \pi(x) I_{b_x} K_{b_x}(X-x)dx - \pi(X)\right)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\frac{(1-R)\ell(Z)}{1-Q_0}\right\|^4\right]^{1/2} \cdot \mathbb{E}\left[\left\|\left(\int_{I_x} \pi(X-b_x u)K_x(u)du - \pi(X)\right)\right\|^4\right]^{1/2} \\
&= O(1)o(1) = o(1),
\end{aligned}
$$

where the second to last equality follows by Assumption 2.1(d), Assumption 2.7(e)(iii), and by DCT. This concludes the derivation of the linear expansion for the second term.

Linear expansions for terms involving ρ_1 and ρ_3 follow by standard (functional) delta method. Hence, we provide the influence functions directly. Let

$$\psi_{p,x,1}(a;\theta_0,G) \equiv \mathbb{E}\left[\frac{(1-R)\ell(Z)\Lambda_x(W;\beta_0)}{(1-Q_0)f_{X|R}(X|1)} \cdot \left(\frac{(1-r)\ell(z)1(x \le X)}{1-Q_0} - F_{X|R}(X|1)\right)\right], \tag{2.8.25}$$

$$\psi_{p,x,3}(a;\theta_0,G) \equiv -\frac{r-Q_0}{Q_0} \cdot \mathbb{E}\left[\frac{(1-R)\ell(Z)\Lambda_x(W;\beta_0)}{1-Q_0} \cdot \frac{G(X)}{f_{X|R}(X|1)}\right]. \tag{2.8.26}$$

Combining (2.8.25), (2.8.23), and (2.8.26), we deduce that

$$\Delta_{p,3} = \frac{1}{n}\sum_{i=1}^{n} \psi_{p_3,i} = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{f_{Y|R}(q_\tau|1)}\left(\psi_{g,p,i} + \frac{(1-R_i)\ell(Z_i)\Lambda_x(W_i;\beta_0)g_p(X_i)}{1-Q_0}\right.$$
$$\left. - d_p(\theta_0,G)\right) + o_p(n^{-1/2}). \tag{2.8.27}$$

where $\psi_{g,p}$ is defined in (2.7.10). Collecting the results in (2.8.20), (2.8.21), and (2.8.27), it follows that the asymptotic linear representation for $\widehat{UQE}_p$ is given as in (2.7.3). This completes our proof. ■

2.8.2 Auxiliary Lemmas and Proofs

In this section, we present and prove some auxiliary lemmas that facilitate the proofs in the previous section.

Lemma 2.4 Under the assumptions of Lemma 2.2, we have

$$\sup_{\theta\in\Theta} ||\mathbb{E}_n[m_4(A;\theta,\widehat{q}_\tau) - m_4(A;\theta,q_\tau)]|| = O_p(n_s^{-1/2}).$$

Proof of Lemma 2.4: To establish the claim, we will first show that

$$\sup_{\theta\in\Theta} |\mathbb{E}_n[\Delta m_4(A;\theta,\widehat{q}_\tau,q_\tau)]| = o_p(n_s^{-1/2}), \tag{2.8.28}$$

where $\Delta m_4(a;\theta,\widehat{q}_\tau,q_\tau) \equiv m_4(a;\theta,\widehat{q}_\tau) - m_4(a;\theta,q_\tau) - \mathbb{E}[m_4(A;\theta,\widehat{q}_\tau) - m_4(A;\theta,q_\tau)|\mathfrak{Y}_n]$, and $\mathbb{E}[\cdot|\mathfrak{Y}_n]$ is the expectation conditional on $\mathfrak{Y}_n = \{Y_i\}_{i=1}^n$. Define

$$\mathscr{F} \equiv \left\{(r,y,z) \in \{0,1\} \times \mathscr{Y}\mathscr{Z} \mapsto \frac{\widetilde{e}(z)rL(k(z)'\gamma)}{L(k(z)'\gamma + t(z)'\lambda_s)}(1(y \le q) - 1(y \le q_\tau)) :\right.$$
$$\left. \widetilde{e}(z) \in \{e_1(z),...,e_{d_\beta}(z)\}, (\gamma,\lambda_s) \in \Theta_{\gamma,\lambda_s}, q \in \mathscr{Y}\right\},$$

where $e_j(z)$ is the j-th element of $e(z)$. Both $m_4(a;\theta,\widehat{q}_\tau)$ and $m_4(a;\theta,q_\tau)$ belong to $\mathscr{F}$.

The result is obtained by Theorem 2.1 in Wellner and van der Vaart (2007). To invoke the theorem, we need to verify two conditions: (i) the class $\mathscr{F}$ is Donsker, and (ii) the variance of $\Delta m_4(A;\theta,\widehat{q}_\tau,q_\tau)$ tends to zero.

To show (i), we let

$$\begin{aligned}
&\mathscr{F}_1 \equiv \{y \in \mathscr{Y} \mapsto 1(y \leq q) - 1(y \leq q_\tau) :\ q \in \mathscr{Y}\},\\
&\mathscr{F}_2 \equiv \{z \in \mathscr{Z} \mapsto L(k(z)'\gamma)/L(k(z)'\gamma + t(z)'\lambda_s) :\ (\gamma,\lambda_s) \in \Theta_{\gamma,\lambda_s}\},\\
&\mathscr{F}_3 \equiv \{(r,z) \in \{0,1\} \times \mathscr{Z} \mapsto re_j(z) :\ j = 1,...,d_\beta\}.
\end{aligned}$$

Hence, $\mathscr{F} \subseteq \mathscr{F}_1 \cdot \mathscr{F}_2 \cdot \mathscr{F}_3$.

Here, $\mathscr{F}_1$ is a collection of indicator functions over $\mathbb{R}$. Under Assumptions 2.7(a)(ii) and (e)(i) , $\mathscr{F}_2$ is pointwise compact; for definition, see Example 19.8 of Van der Vaart (1998). It is well-known that the pointwise compact class of functions and collections of indicators of cells in Euclidean space are pointwise measurable; see the discussion in Section 2.3 of Van Der Vaart and Wellner (1996) for definition. The pointwise measurability of $\mathscr{F}_3$ follows by definition. Then, by Lemma 8.10 in Kosorok (2008), $\mathscr{F}$ is pointwise measurable. $\mathscr{F}_1$ is a VC-subgraph class with VC-index equal to 2. In addition, $\mathscr{F}_1$ is uniformly bounded by $F_1 \equiv 2$. Theorem 2.6.7 in Van Der Vaart and Wellner (1996) implies that

$$\sup_Q N(\varepsilon \|F_1\|_{Q,2}, \mathscr{F}_1, L_2(Q)) \leq K_1 \varepsilon^{-2}. \tag{2.8.29}$$

Due to Assumption 2.7(e)(ii), $\left|\frac{L(k(z)'\gamma_1)}{L(k(z)'\gamma_1+t(z)'\lambda_{s1})} - \frac{L(k(z)'\gamma_2)}{L(k(z)'\gamma_2+t(z)'\lambda_{s2})}\right| \leq K_2 \cdot \|k(z)\|\,\|\gamma_1 - \gamma_2\| + K_3 \cdot \|t(z)\|\,\|\lambda_{s1} - \lambda_{s2}\|$. Combining this fact and Assumption 2.7(f), $\mathscr{F}_2$ admits an integrable envelop function, $F_2 \equiv K_2\bar{\gamma}\|k(z)\| + K_3\bar{\lambda}_s\|t(z)\| + K_4$, where $\bar{\gamma} \equiv \sup_{\gamma\in\Theta_\gamma}\|\gamma\|$ and $\bar{\lambda}_s \equiv \sup_{\lambda_s\in\Theta_{\lambda_s}}\|\lambda_s\|$ are finite under Assumption 2.7(a)(ii). It then follows from Theorem 2.10.20 of Van Der Vaart and Wellner (1996) that for all $\delta > 0$,

$$\begin{aligned}
\int_0^\delta \sup_Q \sqrt{\log N(\varepsilon \|F_2\|_{Q,2}, \mathscr{F}_2, L_2(Q))}d\varepsilon &\leq \sum_{j\in\{\gamma,\lambda_s\}} \int_0^\delta \sqrt{\log N(\varepsilon \bar{j}, \Theta_j, \|\cdot\|)}d\varepsilon\\
&\leq \sum_{j\in\{\gamma,\lambda_s\}} \sqrt{d_j}\int_0^\delta \sqrt{\log\left(1+\frac{4diam(\Theta_j)}{\varepsilon\bar{j}}\right)}d\varepsilon < \infty,
\end{aligned} \tag{2.8.30}$$

where $diam(\Theta)$ is the diameter of Θ, and Q is any finite discrete measure. By Assumption 2.7(f), $\mathbb{E}[F_2^2] < \infty$, and therefore, $\mathscr{F}_2$ is Donsker. Another application of Theorem 2.10.20 of Van Der Vaart and Wellner (1996) to $\mathscr{F}$ yields that

$$\int_0^\delta \sup_Q \sqrt{\log N(4\varepsilon \|F\|_{Q,2}, \mathscr{F}, L_2(Q))}d\varepsilon$$

$$\leq \sum_{j=1}^{3} \int_{0}^{\delta} \sup_{Q} \sqrt{\log N(\varepsilon \left\|F_j\right\|_{Q,2}, \mathscr{F}_j, L_2(Q))} d\varepsilon < \infty, \tag{2.8.31}$$

where $F_3(z) \equiv \|e(z)\|$, is an envelop function for $\mathscr{F}_3$. The last inequality follows from (2.8.29), (2.8.30), and the fact that $\mathscr{F}_3$ is a finite set, and therefore, BUEI.

To show that F is square integrable, note

$$\mathbb{E}[|F_1F_2F_3|^2] \leq 4\mathbb{E}[\|e(Z)\|^2 (K_2\bar{\gamma}\|k(Z)\| + K_3\bar{\lambda}_s \|t(Z)\| + K_4)^2] < \infty,$$

where the second inequality follows by the Cauchy–Schwartz inequality and Assumption 2.7(f). By Theorem 2.10.20 and Theorem 2.10.1 in Van Der Vaart and Wellner (1996), $\mathscr{F}$ is Donsker.

In the next step, we show (ii). That is, $\int \|\Delta m_4(a;\theta,\widehat{q}_\tau,q_\tau)\|^2 dF_A(a)$ converges in probability to 0. Towards this end,

$$\begin{aligned}\int ||\Delta m_4(a;\theta,\widehat{q}_\tau,q_\tau)||^2 dF(a) &\leq \frac{c_5^2}{2c_4^2} \cdot \left(\sup_{y\in\mathscr{Y},z\in\mathscr{Z}} |f_{Y|ZR}(y|z,1)|\right)^2 \mathbb{E}[\|e(Z)\|^2](\widehat{q}_\tau - q_\tau)^2 \\ &\leq K_5(\widehat{q}_\tau - q_\tau)^2 = O_p(n_s^{-1}),\end{aligned} \tag{2.8.32}$$

where the second inequality is by MVT and Assumption 2.7(b). The last equality follows by (2.8.7).

Combining (i) and (ii), Theorem 2.1 in Wellner and van der Vaart (2007) yields (2.8.28). To complete the proof, it suffices to show that $\mathbb{E}[m_4(A;\theta,\widehat{q}_\tau) - m_4(A;\theta,q_\tau)|\mathfrak{Y}_n] = O_p(n^{-1/2})$. By a first-order expansion,

$$\begin{aligned}&\mathbb{E}[\|m_4(A;\theta,\widehat{q}_\tau) - m_4(A;\theta,q_\tau)\| \mid \mathfrak{Y}_n] = \mathbb{E}[\left\|Rf_{Y|ZR}(\widetilde{q}_\tau|Z,1)e(Z)\right\| \mid \mathfrak{Y}_n] \cdot |\widehat{q}_\tau - q_\tau| \\ &\leq K_6 \sup_{y\in\mathscr{Y},z\in\mathscr{Z}} |f_{Y|ZR}(y|Z,1)|\mathbb{E}[\|e(Z)\|] \cdot O_p(n^{-1/2}) = O_p(1)\cdot O_p(n^{-1/2}) = O_p(n^{-1/2}),\end{aligned}$$

where the second line follows by MVT, Assumptions 2.7(b), and (f). ■

Lemma 2.5 Under the assumptions of Lemma 2.2,

$$\begin{aligned}\mathscr{H}_1 &\equiv \left\{(r,z) \in \{0,1\} \times \mathscr{Z} \mapsto \frac{(r - L(k(z)'\gamma))L'(k(z)'\gamma)\widetilde{k}(z)}{L(k(z)'\gamma)(1 - L(k(z)'\gamma))} : \widetilde{k}(z) \in \{k_1(z),...,k_{d_\gamma}(z)\}, \gamma \in \Theta_\gamma\right\}, \\ \mathscr{H}_2 &\equiv \left\{(r,z) \in \{0,1\} \times \mathscr{Z} \mapsto \left(\frac{r}{L(k(z)'\gamma + t(z)'\lambda_s)} - 1\right) L(k(z)'\gamma)\widetilde{t}(z) : \right. \\ &\qquad\qquad \left. \widetilde{t}(z) \in \{t_1(z),...,t_{d_\lambda}(z)\}, (\gamma,\lambda_s) \in \Theta_{\gamma,\lambda_s}\right\},\end{aligned}$$

$$\mathscr{H}_3 \equiv \left\{ (r,z) \in \{0,1\} \times \mathscr{Z} \mapsto \left(\frac{1-r}{1-L(k(z)'\gamma + t(z)'\lambda_a)} - 1 \right) L(k(z)'\gamma)\widetilde{t}(z) : \right.$$

$$\left. \widetilde{t}(z) \in \{t_1(z),...,t_{d_\lambda}(z)\}, (\gamma,\lambda_a) \in \Theta_{\gamma,\lambda_a} \right\},$$

$$\mathscr{H}_4 \equiv \left\{ (r,z,x,y) \in \{0,1\} \times \mathscr{Z}\mathscr{X}\mathscr{Y} \mapsto \left(\frac{r \cdot 1(y \le q)}{L(k(z)'\gamma + t(z)'\lambda_s)} - \frac{(1-r)\cdot\Lambda(w,\beta)}{1-L(k(z)'\gamma + t(z)'\lambda_a)} \right) \cdot L(k(z)'\gamma)\widetilde{e}(z) : \right.$$

$$\left. \widetilde{e}(z) \in \{e_1(z),...,e_{d_\beta}(z)\}, q \in \mathscr{Y}_0, \theta \in \Theta \right\},$$

are Glivenko-Cantelli (GC).

Proof of Lemma 2.5: Under Assumptions 2.7(e)(iii) and (f), $\mathscr{H}_1, \mathscr{H}_2$, and $\mathscr{H}_3$, admit integrable envelop functions, H_1, H_2, and H_3, respectively, where $H_1(r,z) \equiv K_1 \|k(z)\|$, $H_2(r,z) \equiv K_2 \|t(z)\|$, and $H_3(r,z) \equiv K_3 \|t(z)\|$. Hence, by Example 19.8 in Van der Vaart (1998), $\mathscr{H}_j$, $j = 1,2,3$, are GC. To show that $\mathscr{H}_4$ is also GC, we define

$$\mathscr{H}_5 \equiv \{y \in \{0,1\} \times \mathscr{Y} \mapsto 1(y \le q), q \in \mathscr{Y}_0\},$$

$$\mathscr{H}_6 \equiv \left\{ (r,z,x) \in \{0,1\} \times \mathscr{Z}\mathscr{X}\mathscr{Y} \mapsto \frac{(1-r)\Lambda(w,\beta)L(k(z)'\gamma)\widetilde{e}(z)}{1-L(k(z)'\gamma + t(z)'\lambda_a)}, \right.$$

$$\left. \widetilde{e}(z) \in \{e_1(z),...,e_{d_\beta}(z)\}, q \in \mathscr{Y}_0, \theta \in \Theta \right\}.$$

Notice that $\mathscr{H}_4 = \mathscr{F}_2\mathscr{F}_3\mathscr{H}_5 - \mathscr{H}_6$. Under Assumptions 2.7(c)(iii), (e)(iii), and (f), $\mathscr{H}_6$ admits an integrable envelop, $H_6 \equiv K_4 \|e(z)\|$. By Example 19.8 in Van der Vaart (1998), $\mathscr{H}_4$ is GC. It is straightforward to check $\mathscr{F}_2\mathscr{F}_3\mathscr{H}_5$ admits an integrable envelop function. From the proof of Lemma 2.4, we know that $\mathscr{F}_2$ and $\mathscr{F}_3$ are Donsker. Then, by the Donskerness of $\mathscr{H}_5$ and $\mathscr{H}_6$, and Corollary 9.26 in Kosorok (2008), $\mathscr{H}_4$ is also GC. ■

Lemma 2.6 Under the assumptions of Theorem 2.3,

$$\widehat{d_{q,n}}(\widehat{\theta}, G) - d_q(\theta_0, G) = o_p(1).$$

Proof of Lemma 2.6: Let $\eta(\theta) = (\eta_1(\cdot,\theta), \eta_2)'$, $f(a;\theta,\eta(\theta)) \equiv f_1(a;\theta,\eta(\theta)) \cdot G^{-1}(\eta_1(x,\theta)/\eta_2) + f_2(a;\theta,\eta(\theta))$, where

$$f_1(a;\theta,\eta(\theta)) \equiv \frac{1-r}{Q_0\eta_2} \cdot \frac{L(k(z)'\gamma + t(z)'\lambda_s)}{1-L(k(z)'\gamma + t(z)'\lambda_a)} \cdot \Lambda_x(w;\beta),$$

and $f_2(a;\eta(\theta)) \equiv -x \cdot f_1(a;\theta,\eta(\theta))$. With these quantities, we can write

$$\widehat{d_{q,n}}(\widehat{\theta}, G) - d_q(\theta_0, G) = \mathbb{E}_n[f(A;\theta_0,\widehat{\eta}_n(\theta))] - \mathbb{E}[f(A;\eta_0(\theta_0))],$$

where

$$\widehat{\eta}_n(\theta) \equiv \left(\mathbb{E}_n \left[\frac{(1-R)}{Q_0} \frac{L(k(z)'\widehat{\gamma}+t(z)'\widehat{\lambda}_s)}{1-L(k(z)'\widehat{\gamma}+t(z)'\widehat{\lambda}_a)} 1(X \leq \cdot) \right], \frac{\mathbb{E}_n[R]}{Q_0} \right)',$$
$$\eta_0(\theta) \equiv \left(\mathbb{E} \left[\frac{(1-R)}{Q_0} \frac{L(k(z)'\gamma+t(z)'\lambda_s)}{1-L(k(z)'\gamma+t(z)'\lambda_a)} 1(X \leq x) \right], 1 \right)'.$$

Let $\mathscr{N}_{\theta_0}$ denote a neighborhood of θ_0. We proceed by showing, (i) when $\widehat{\eta}_n(\theta)$ is sufficiently close to $\eta_0(\theta)$, we have $\sup_{\theta \in \mathscr{N}_{\theta_0}} \|f(a;\theta,\eta(\theta)) - f(a;\theta,\eta_0(\theta))\| \leq c(a) \cdot \sup_{x\in\mathscr{X},\theta\in\mathscr{N}_{\theta_0}} \|\eta(\theta) - \eta_0(\theta)\|$, for some $c(\cdot)$ satisfying, (ii)

$$\mathbb{E}[c(A)] \sup_{x\in\mathscr{X},\theta\in\mathscr{N}_{\theta_0}} \|\widehat{\eta}_n(\theta) - \eta_0(\theta)\| \xrightarrow{p} 0,$$

and (iii) $\mathbb{E}[\sup_{\theta\in\mathscr{N}_{\theta_0}, \|\eta-\eta_0\|\leq\varepsilon} \|f(A;\theta,\eta(\theta))\|] < \infty$.

By Assumptions 2.7(c) and (e), $L(\cdot)$ and $\Lambda_x(\cdot)$ are uniformly bounded. Under Assumption 2.8(d), G^{-1} is Lipschitz with a bounded Lipschitz constant. Then, (i) follows immediately by

$$\|f(a;\theta,\eta(\theta)) - f(a;\theta,\eta_0(\theta))\| \leq K_1 \sup_{x\in\mathscr{X}} \|\eta_1(x,\theta) - \eta_{1,0}(x,\theta)\| + K_2 \|\eta_2 - \eta_{2,0}\|,$$

and letting $c(\cdot) = \max\{K_1, K_2\}$.

Next, we shall verify (ii). To this end, it suffices to show that $\widehat{\eta}_{1,n}$ converges to $\eta_{1,0}$ uniformly. This follows from $\mathscr{G}_\eta$ being GC, where $\mathscr{G}_\eta \equiv \mathscr{G}_{\eta,1} \cdot \mathscr{G}_{\eta,2}$, and

$$\mathscr{G}_{\eta,1} \equiv \left\{ (r,z) \in \{0,1\} \times \mathscr{X} \mapsto \frac{(1-r)L(k(z)'\gamma+t(z)'\lambda_s)}{Q_0(1-L(k(z)'\gamma+t(z)'\lambda_a))} : \theta \in \Theta \right\},$$
$$\mathscr{G}_{\eta,2} \equiv \{x \in \mathscr{X} \mapsto 1(x \leq q) : q \in \mathscr{X}\}.$$

Since $\mathscr{G}_{\eta,1}$ is uniformly bounded, it follows from Example 19.2 in Van der Vaart (1998) that it is GC. $\mathscr{G}_{\eta,2}$ is VC subgraph with VC index equal to 2, and therefore, it is GC. Then, by Corollary 9.27 in Kosorok (2008), $\mathscr{G}_\eta$ is also uniformly bounded GC.

Lastly, (iii) holds under Assumption 2.1(d) and Assumption 2.7(e).

By (i), (ii), and the Markov inequality, $\sup_{\theta\in\mathscr{N}_{\theta_0}} \|f(A;\theta,\widehat{\eta}(\theta)) - f(A;\theta,\eta_0(\theta))\| \xrightarrow{p} 0$. Then the desired result is obtained by (iii) and Lemma 4.3 in Newey and McFadden (1994). ■

Lemma 2.7 Under the assumptions of Theorem 2.3, we have that $\mathbb{E}_n[\Delta\phi_q(A;\theta_0,G)] = n^{-1}\sum_{i=1}^n \psi_{g,q}(A_i;\theta_0,G) + o_p$ $(n^{-1/2})$, where $\Delta\phi_g$ and $\psi_{g,q}$ are defined in (2.8.17) and (2.7.9), respectively.

Proof of Lemma 2.7: Let $f_0(a;\eta) \equiv \frac{1-r}{1-Q_0}\ell(z)\Lambda_x(w;\beta_0)G^{-1}(\eta_1/\eta_2)$, and hence, $\Delta\phi_q(a;\theta_0,G) = f_0(a;\widehat{\eta}_n) - f_0(a;\eta_0)$, where $\widehat{\eta}_n(\cdot) \equiv \left(\mathbb{E}_n\left[\frac{(1-R)\ell(Z)}{1-Q_0}1(X \leq \cdot)\right], \mathbb{E}_n[R]/Q_0\right)'$, and $\eta_0(\cdot) \equiv (F_{X|R}(\cdot|1),\ 1)'$. The first step of our proof is to show that

$$\sqrt{n}(\mathbb{E}_n - \mathbb{E})(f_0(a;\widehat{\eta}_n) - f_0(a;\eta_0)) = o_p(1), \tag{2.8.33}$$

by invoking Theorem 2.1 in Wellner and van der Vaart (2007).

Towards this end, we need to (i) define the functional space $\mathscr{H}_\eta$ such that $\mathbb{P}(\widehat{\eta}_n \in \mathscr{H}_\eta) \to 1$, (ii) verify that $\mathscr{F}_\eta \equiv \{f_0(\cdot;\eta) : \eta \in \mathscr{H}_\eta\}$ is Donsker, and (iii) show that

$$\int (f_0(a;\widehat{\eta}) - f_0(a;\eta_0))^2 dF_A(a) \xrightarrow{p} 0. \tag{2.8.34}$$

For (i), let $\mathscr{H}_\eta \equiv (\mathscr{H}_{\eta,1}, \mathscr{H}_{\eta,2})$, where

$$\mathscr{H}_{\eta,1} \equiv \{x \in \mathscr{X} \mapsto f(x) : f \text{ non-decreasing, bounded between 0 and 1}\},$$

and $\mathscr{H}_{\eta,2} \equiv [1-\delta_\eta, 1+\delta_\eta]$, for some $\delta_\eta \in (0,1/2)$. Given $\mathscr{H}$, Condition (i) is implied by $\widehat{\eta} \xrightarrow{p} \eta_0$ uniformly, which is shown in Lemma 2.6.

Next, we establish the Donsker property of $\mathscr{F}_\eta$. Note that $\mathscr{F}_\eta = \frac{1-r}{1-Q_0}\ell(z)\Lambda_x(w;\beta_0)\cdot G^{-1}(\mathscr{H}_{\eta_1}/\mathscr{H}_{\eta_2})$. By Lemma 9.11 in Kosorok (2008), $\mathscr{H}_{\eta,1}$ is BUEI, relative to the envelop $H_{\eta,1} \equiv 1$. $\mathscr{H}_2$ is a bounded convex set in the Euclidean space, and hence, trivially BUEI. Pointwise measurability is immediate from the definitions of the two sets. By Assumption 2.8(d), G^{-1} is a Lipschitz continuous function with a bounded Lipschitz constant. Given that $\mathscr{H}_2$ is bounded away from 0, we conclude from Lemma 9.14 and Theorem 9.15 in Kosorok (2008) that $G^{-1}(\mathscr{H}_{\eta_1}/\mathscr{H}_{\eta_2})$ is also BUEI and pointwise measurable relative to the envelop $H_{\eta,3} \equiv \sup\{x \in \mathscr{X}\}$. By Theorem 2.10.1 in Van Der Vaart and Wellner (1996), $G^{-1}(\mathscr{H}_{\eta_1}/\mathscr{H}_{\eta_2})$ is uniformly bounded Donsker. Finally, under Assumptions 2.7(c)(iii) and (e), Corollary 9.32 then implies that $\mathscr{F}_\eta$ is uniformly bounded Donsker.

Now, to show (2.8.34), note that

$$\int (f_0(a;\widehat{\eta}) - f_0(a;\eta_0))^2 dF_A(a) \leq K_1 \sup_{x\in\mathscr{X}} |\widehat{\eta}_{1,n}(x) - \eta_{1,0}(x)|^2 + K_2|\widehat{\eta}_{2,n} - \eta_{2,0}|^2 = o_p(1),$$

where the first inequality follows by carefully bounding the coefficients associated with each term and by the fact that fourth moments of $\|k(z)\|$ and $\|t(z)\|$ exist under Assumption 2.7(f). The last one follows because $\widehat{\eta}_n$ converges uniformly to η_0.

Next, we prove that

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\eta}_{1,n}(\cdot)\\ \widehat{\eta}_{2,n}\end{pmatrix}-\begin{pmatrix}\eta_{1,0}(\cdot)\\ \eta_{2,0}\end{pmatrix}\right)\rightsquigarrow\mathbb{G}_{\eta_1,\eta_2}(\cdot),$$

where $\mathbb{G}_{\eta_1,\eta_2}(\cdot)$ is a tight, two-dimensional mean zero Gaussian process with covariance function $\Sigma(x_1,x_2)=\mathbb{E}[(\psi_{\eta,1}(x_1),\psi_{\eta,2})(\psi_{\eta,1}(x_2),\psi_{\eta,2})']$, $\psi_{\eta,1}(\cdot)=\psi_{\eta,1}(a,\gamma_0;\cdot)\equiv h_\eta(r,z;\gamma_0)1(x\leq\cdot)-F_{X|R}(x|1)$, $\psi_{\eta,2}\equiv(r-Q_0)/Q_0$, and $h_\eta(r,z;\gamma_0)\equiv\frac{(1-r)\ell(z)}{1-Q_0}$.

The weak convergence follows from $\mathscr{G}_{0,\eta}$ being Donsker, where

$$\mathscr{G}_{0,\eta}\equiv\{(r,x,z)\in\{0,1\}\times\mathscr{X}\times\mathscr{Z}\mapsto h_\eta(r,z;\gamma_0)\cdot 1(x\leq q):q\in\mathscr{X}\}.$$

Under Assumption 2.7(e), h_η is uniformly bounded. Since the class of indicator functions is uniformly bounded Donsker, by Corollary 9.32 in Kosorok (2008), we conclude that $\mathscr{G}_\eta$ is also uniformly bounded Donsker.

In view of (2.8.33), we deduce that

$$\mathbb{E}_n[\Delta\phi_q(A;\theta_0,G)]=\mathbb{E}[f_0(A;\widehat{\eta}_n)-f_0(A;\eta_0)]+o_p(n^{-1/2}).\tag{2.8.35}$$

In the next step, we derive the asymptotic linear representation of the first term on the right hand side of (2.8.35) . The proof continues by showing that the map, $\phi_G(\eta)\equiv\int f(a;\eta)dF_A(a)$, is Hadamard differentiable in η at η_0. Observe that we can decompose ϕ_G as follows

$$(\eta_1,\eta_2)\mapsto\frac{\eta_1}{\eta_2}\mapsto G^{-1}\circ(\eta_1/\eta_2)\mapsto h_G\circ(G^{-1}\circ(\eta_1/\eta_2)),$$

where $h_G(g)\equiv\int\frac{1-r}{1-Q_0}\ell(z)\Lambda_x(w;\beta_0)gdF_A(a)$. The first map is continuous and uniformly bounded on $\mathscr{H}_{\eta,2}$, and thus, Hadamard differentiable. The second and third maps are both composition maps. Since $G^{-1}(\cdot)$ is continuously differentiable with bounded first-order derivative under Assumption 2.8(d), by Lemma 3.9.25 in Van Der Vaart and Wellner (1996), it is Hadamard differentiable. Since integration is a linear functional, h_G is also Hadamard differentiable. Now we invoke the chain rule, e.g. Theorem 20.9 in Van der Vaart (1998), and conclude that ϕ_G is Hadamard differentiable, with the functional derivative of ϕ_G at η_0 in the direction of $(\psi_{\eta,1},\psi_{\eta,2})$ given as follows

$$(\psi_{\eta,1},\psi_{\eta,2})\mapsto h_g\circ\left(\left(\frac{1}{G'}\circ G^{-1}\circ\left(\frac{\eta_{1,0}}{\eta_{2,0}}\right)\right)\cdot\left(\frac{\psi_{\eta,1}}{\eta_{2,0}}-\frac{\eta_1\psi_{\eta,2}}{\eta_{2,0}^2}\right)\right).$$

Given the Hadamard differentiability, we then apply the delta method (as in Theorem 3.9.4 in Van Der Vaart and

Wellner (1996)) to get the linear expansion, $\psi_{g,q}$,

$$\begin{aligned}\psi_{g,q}(a;\theta_0) &= \int \frac{1-r}{1-Q_0} \cdot \frac{\ell(z)\Lambda_x(w;\beta_0)}{G'\left(G^{-1}(\eta_{1,0}/\eta_{2,0})\right)} \cdot \left(\frac{\psi_{\eta,1}}{\eta_{2,0}} - \frac{\eta_{1,0}\psi_{\eta,2}}{\eta_{2,0}^2} \right) dF_A(a) \\ &= \mathbb{E}\left[\frac{1-R}{1-Q_0} \cdot \frac{\ell(Z)\Lambda_x(W;\beta_0)}{G'\left(G^{-1}(F_{X|R}(X|1))\right)} \cdot \left(\frac{(1-r)\ell(z)1(x\leq X)}{1-Q_0} - \frac{rF_{X|R}(X|1)}{Q_0} \right) \right].\end{aligned}$$

This concludes our proof. ■

2.8.3 Asymptotic Variance Estimators

In this section, we provide an estimator for the improved asymptotic variance $\widehat{\Sigma}_{j,imp}$ via plug-in estimators of the influence functions $\widehat{\psi}_j$, $j=p,q$. In view of (2.7.3)–(2.7.5) , we define, for $j=p,q$,

$$\widehat{\psi}_j(a;\widehat{\theta},\widehat{q}_\tau,b_y) \equiv \widehat{\psi}_{f_y,j}(a;\widehat{\theta},\widehat{q}_\tau,b_y) - \frac{1}{\widehat{f}_{Y|R}(\widehat{q}_\tau|1)}\widehat{\psi}_{d,j}(a;\widehat{\theta},\widehat{q}_\tau,b_y), \tag{2.8.36}$$

where

$$\widehat{\psi}_{f_y,j}(a;\widehat{\theta},\widehat{q}_\tau,b_y) \equiv \frac{\widehat{d}_{j,n}(\widehat{\theta},G)}{\widehat{f}^2_{Y|R}(\widehat{q}_\tau|1)} \frac{r}{\mathbb{E}_n[R]} \left(K_{b_y}(y-\widehat{q}_\tau) - \mathbb{E}_{n_s}[K_{b_y}(y-\widehat{q}_\tau)] - \frac{(1(y\leq \widehat{q}_\tau)-\tau)\widehat{f}'_{Y|R}(\widehat{q}_\tau|1)}{\widehat{f}_{Y|R}(\widehat{q}_\tau|1)} \right),$$

and

$$\widehat{\psi}_{d,j}(a;\widehat{\theta},\widehat{q}_\tau) \equiv \widehat{M}_{\theta,j,n}(\widehat{\theta})'\widehat{\psi}_\theta(a;\widehat{\theta},\widehat{q}_\tau) + \widehat{\psi}_{g,j}(a;\widehat{\theta},G) + \frac{1-r}{\mathbb{E}_n[1-R]}\widehat{\ell}(z)\Lambda_x(w;\widehat{\beta})\widehat{g}_j(x) - \frac{rd_{j,n}(\widehat{\beta},G)}{\mathbb{E}_n[R]}.$$

First, we focus on $\widehat{\psi}_{f_y}$. The only term that requires some explanation here is $\widehat{f}'_{Y|R}(\widehat{q}_\tau|1)$. We can estimate $\widehat{f}'_{Y|R}$ $(\widehat{q}_\tau|1)$ by $\mathbb{E}_{n_s}[K'_{b_y}(y-\widehat{q}_\tau)]$, where $K'_{b_y}(\cdot) \equiv b_y^{-1}\partial K_y(u)/\partial u$, is the first-order derivative of the rescaled kernel function.

Second, we turn to the influence functions involving the estimation of θ. We let

$$\widehat{\psi}_\theta(a;\widehat{\theta},\widehat{q}_\tau) \equiv -\left(\widehat{M}'\widetilde{\Omega}_n\widehat{M}\right)^{-1}\widehat{M}'\widetilde{\Omega}_n\left(m(a;\widehat{\theta},\widehat{q}_\tau) - \widehat{M}_{q_\tau}\frac{r\cdot(1(y\leq\widehat{q}_\tau)-\tau)}{\mathbb{E}_n[R]\widehat{f}_{Y|R}(\widehat{q}_\tau|1)} \right),$$

where $\widehat{M} \equiv \mathbb{E}_n[\partial m(A;\widehat{\theta},\widehat{q}_\tau)/\partial\theta]$, $\widehat{M}_{q_\tau} \equiv \left(0',0',0',\mathbb{E}_n[Re(Z)\widehat{f}_{Y|R,Z}(\widehat{q}_\tau|1,Z)]'\right)'$, and

$$\widehat{f}_{Y|R,Z}(\widehat{q}_\tau|1,z) \equiv \frac{\mathbb{E}_{n_s}[K_{b_z}(Z-z)K_{b_y}(Y-\widehat{q}_\tau)]}{\mathbb{E}_{n_s}[K_{b_z}(Z-z)]},$$

for $K_{b_z}(Z-z) \equiv \prod_{k=1}^{d_z}\frac{1}{b_{z_k}}K_z\left(\frac{Z_k-z_k}{b_{z_k}}\right)$.

Lastly, we provide estimators for influence functions relating to the estimation of nonparametric first steps. The estimation of $\psi_{g,q}(a;\widehat{\theta},G)$ is straightforward. We let $\widehat{\psi}_{g,p}(a;\widehat{\theta},G) \equiv \sum_{j=1}^{3} \widehat{\psi}_{p,x,j}(a;\widehat{\theta},G)$. The estimation of $\psi_{p,x,2}$ calls for some explanation. Note that the plug-in estimator,

$$\widehat{\psi}_{p,x,2}(a;\widehat{\theta},G) \equiv \frac{(1-r)\widehat{\ell}(z)}{\mathbb{E}_n[1-R]}\widehat{\pi}(x) - \mathbb{E}_{n_a}\left[\widehat{\ell}(z)\widehat{\pi}(x)\right],$$

depends on an estimator for $\pi(\cdot)$. To estimate $\pi(\cdot)$, we rewrite it as follows

$$\pi(x) = \mathbb{E}\left[\left.\frac{\ell(Z)\Lambda_x(W)(G(X)-F_{X|R}(X|1))}{f_{X|R}(X|1)}\right| X=x, R=0\right].$$

Let $\phi_\pi(\cdot)$ denote the term inside the conditional expectation operator, and $\widehat{\phi}_\pi(\cdot)$ be its sample analog. Here we propose a simple series least square estimator for $\pi(\cdot)$. Let $P_k(x) = (p_1(x),...,p_k(x))'$ be a k-dimensional vector of known basis functions with $k \to \infty$ and $nk \to 0$. Then, we can estimate $\pi(x)$ by letting $\widehat{\pi}(x) \equiv P_k(x)'\mathbb{E}_{n_a}[P_k(X)P_k(X)']^{-1}$ $\cdot\mathbb{E}_{n_a}[P_k(X)\widehat{\phi}_\pi(X)]$.

CHAPTER 3

Difference-in-Differences with Compositional Changes

This chapter is adapted from the working paper “Difference-in-Differences with Compositional Changes” and has been reproduced with the permission of my co-author Pedro H. C. Sant’Anna.

3.1 Introduction

Difference-in-differences (DiD) designs have been used widely for identifying and estimating causal effects with observational data. Identification in this research design typically relies on a conditional parallel trends assumption stipulating that conditional on a set of covariates, the average untreated outcomes among treated and comparison groups would have evolved “in parallel”. When one pairs this assumption with common support and no-anticipation assumptions, it is easy to establish that the average treatment effect on the treated (ATT) is nonparametrically identified when panel data is available. When one only observes repeated cross-sectional data, it is common to impose further a no-compositional change assumption, also known as the stationarity assumption. This is the case in the widely cited DiD procedures of Heckman et al. (1997), Abadie (2005), Sant’Anna and Zhao (2020), and Callaway and Sant’Anna (2021), for example.

Although we have seen a lot of recent developments in DiD methods (see Roth et al., 2023 for an overview of recent DiD developments), little attention has been paid to understanding the importance and limitations of the no-compositional changes assumption. This paper aims to fill this gap by providing researchers with new tools that can be used when they are in doubt about such an assumption and/or to test its plausibility.

Before discussing the paper’s contributions, it is worth stressing why ruling out compositional changes across time periods can be restrictive in real empirical applications. Essentially, the no-compositional changes assumption requires one to sample observations from the same population across time periods, which can be unrealistic in some scenarios. For example, Hong (2013) studies the effect of Napster on recorded music sales. He uses data from the 1996–2002 Interview Surveys of the Consumer Expenditure Survey. Over this period, the composition of internet users has changed substantially. The early adopters tend to be younger, richer, more educated, and technically savvy, whereas later adopters exhibit a higher diversity level in demographics. If one ignores such imbalances of group composition across time, the (negative) effect of Napster on music sales can be overestimated, as the decrease in the average music expenditure may be attributed to a post-Napster group with more households having low reservation prices for recorded music. Other applications also share this concern, as discussed below and in more detail in Section

3.6. Therefore, having causal inference tools that can assess if the findings are robust against compositional changes in the sample is of practical interest.

We begin our analysis by showing that one can identify the ATT in DiD setups without invoking the no compositional changes assumption. We derive the efficient influence function and the semiparametric efficiency bound for the ATT in this scenario. We then form generic nonparametric estimators built on the efficient influence function that can achieve the semiparametric efficient bound under mild smoothness conditions, a rate doubly-robust (DR) property (Smucler et al., 2019). These results are general and do not rely on a specific choice of estimators for nuisance functions. Nonetheless, they do not help us with practical inference procedures. For that, we use a local polynomial estimator for the outcome-regressions models and the local multinomial logit regression to estimate the generalized propensity score, the latter of which is fairly new in the DiD literature. Importantly, our nonparametric estimators can accommodate both discrete and continuous covariates, and all tuning parameters are selected in a data-driven way via cross-validation.[1] Finally, we show that the estimand proposed by Sant'Anna and Zhao (2020) is no longer DR in this DiD setup with compositional changes. In fact, we show that even when all nuisance functions are correctly specified, the Sant'Anna and Zhao (2020)'s DR DiD estimand does not identify the ATT in this general setup. Overall, this first set of results highlights what is "the best" that one can do in DiD setups with compositional changes.

Next, we tackle the problem of how much efficiency one may lose by not exploring the no-compositional change assumption when it is valid. To answer this question, we compare our derived semiparametric efficiency bound that does not impose the no-compositional changes assumption with the semiparametric efficiency bound derived by Sant'Anna and Zhao (2020) that fully exploits it. As expected, the extra layer of robustness comes at the cost of loss of efficiency. Heuristically speaking, the no-compositional change assumption allows one to pool the covariate data from all time periods, substantially increasing the effective sample size and the precision of the DiD estimator compared to the one that does not impose the no-compositional change assumption.

In practice, determining whether compositional changes are a significant concern for a given empirical application is not always obvious. Specifically, it is unclear whether imposing a no-compositional change assumption will lead to biased ATT estimates. Using our previous results, we propose a nonparametric Hausman (1978)-type test for no-compositional changes. The test compares our nonparametric DiD estimator of the ATT, which is robust against compositional changes, with the nonparametric extension of Sant'Anna and Zhao (2020)'s DR DiD estimator, which assumes no compositional changes. We derive the large sample properties of the proposed test, which shows that it controls size asymptotically and is consistent against a broad set of alternatives.

We demonstrate the practical appeal of our proposed DiD tools through Monte Carlo simulations and an empirical

[1] As a side contribution of this paper, we provide a new result on the uniform expansion of the local (multinomial) logit estimators, which accommodates both continuous and discrete variables. This result may be of independent interest.

application that revisits Sequeira (2016). She leverages a quasi-experimental variation created by a large reduction in the average nominal tariff rate between South Africa and Mozambique in 2008 to study the causal effect of tariff rate reduction on trade costs and corruption behavior using a two-way fixed effects specification with covariates that implicitly imposes a no-compositional changes assumption, among other arguably unnecessary homogeneity assumptions. We use our nonparametric tests to assess the plausibility of the no-compositional changes assumption and fail to reject it at the usual significance levels. Our results support the conclusions by Sequeira (2016) that tariff liberalization decreases corruption, and our DR DiD estimates are similar to those in the original paper.

Related literature: This article belongs to the extensive literature on semiparametric DiD methods. We refer the reader to Roth et al. (2023) for a synthesis of recent advances in the econometrics of DiD. Within this broad literature, the paper closest to ours is Sant'Anna and Zhao (2020), which proposes DR DiD estimators for the ATT and derives semiparametric efficiency bound for such estimators, too. In sharp contrast to us, though, all the results in Sant'Anna and Zhao (2020) rely on a no-compositional change assumption. Thus, our results complement theirs. Furthermore, Sant'Anna and Zhao (2020)'s theoretical results rely on parametric first-step estimators, while we accommodate nonparametric estimators. A perhaps side and minor contribution of our paper is establishing the statistical properties of Sant'Anna and Zhao (2020)'s DR DiD estimator with nonparametric estimates of the nuisance functions; see also Chang (2020).

Our paper also relates to the causal inference literature on compositional changes over time. Hong (2013) develops a matching-based estimator under a "selection-on-observable"-type assumption, which is different and arguably stronger than our conditional parallel trends assumption. Hong (2013) also does not discuss efficiency issues as we do. Stuart et al. (2014) propose inverse probability weighted estimators for the ATT in DiD setups under compositional changes. In contrast to us, their estimator does not enjoy any DR property and may not attain the semiparametric efficiency bound. Nie et al. (2019) is also interested in DiD estimators under compositional changes. Their estimator substantially differs from ours: they use meta-learners and cross-fitting to estimate nuisance functions, while our estimator is based on the efficient influence function for the ATT. When treatment effects are heterogeneous, their estimators do not target the ATT but the ATE, which, in our context, is not identified. They do not consider tests for the no-compositional changes assumption as we do.

Finally, we contribute to the semiparametric two-stage estimation that depends on nonparametrically estimated functions. See, e.g., Newey (1994), Chen et al. (2003), Chen et al. (2008), Ackerberg et al. (2014), Rothe and Firpo (2019), among many others. Our results on local multinomial logit regression builds on Fan et al. (1995), Claeskens and Van Keilegom (2003), Li and Ouyang (2005), and Kong et al. (2010). The novel result on the uniform expansion of the local multinomial logit estimator may be of independent interest.

Organization of the paper: Section 3.2 introduces the identification framework of the DiD parameter under compositional changes, presents the semiparametric efficiency results, and discusses the bias-variance trade-off of ruling out compositional changes. In Section 3.3, we present our nonparametric DR DiD estimators, discuss their large sample properties, and how to pick tuning parameters. Section 3.4 discusses a test for no-compositional changes. Monte Carlo simulations are provided in Section 3.5, and an empirical illustration is considered in Section 3.6. Section 3.7 concludes. Proofs and additional results are reported in Section 3.8.

3.2 Difference-in-Differences

3.2.1 Framework

This section describes our setup. We focus on the canonical two-period and two-group setup for conciseness and transparency. We have two time periods, $t = 0$, where no unit is exposed to the treatment, and time $t = 1$, where units in the group with $D = 1$ are exposed to treatment; here, D is a binary treatment indicator. We adopt the potential outcome notation where $Y_{it}(0)$ and $Y_{it}(1)$ denote the untreated and treated potential outcome for unit i at time t, respectively. Observed outcomes are given by $Y_{it} = D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0)$. We also assume that a k-dimensional vector of pre-treatment characteristics $X_i \in \mathscr{X} \subseteq \mathbb{R}^k$ is available.

This paper considers the case where one has access to repeated cross-sectional data. To formalize this idea, let T_i be a dummy variable that takes value one if the observation i is observed only in the post-treatment period $t = 1$, and zero if observation i is only observed in the pre-treatment period $t = 0$. Define $Y_i = T_iY_{i1} + (1 - T_i)Y_{i0}$, and let n_1 and n_0 be the sample sizes of the post-treatment and pre-treatment periods such that $n = n_1 + n_0$.

Assumption 3.1 (Sampling) The pooled data $\{Y_i, D_i, X_i, T_i\}_{i=1}^{n}$ consists of independent and identically distributed draws from the mixture distribution

$$\begin{aligned} \mathbb{P}(Y \leq y, D = d, X \leq x, T = t) &= t \cdot \mathbb{P}(T = 1) \cdot \mathbb{P}(Y_1 \leq y, D = d, X \leq x | T = 1) \\ &\quad + (1 - t) \cdot \mathbb{P}(T = 0)\mathbb{P}(Y_0 \leq y, D = d, X \leq x | T = 0), \end{aligned}$$

where $(y, d, x, t) \in \mathscr{Y} \times \{0, 1\} \times \mathscr{X} \times \{0, 1\}$.

Assumption 3.1 allows for different sampling schemes. For instance, it accommodates the binomial sampling scheme where an observation i is randomly drawn from either (Y_1, D, X) or (Y_0, D, X) with a fixed probability. It also accommodates the "conditional" sampling scheme where n_1 observations are sampled from (Y_1, D, X), n_0 observations are sampled from (Y_0, D, X) and $\mathbb{P}(T = 1) = n_1/n$ (here, T is treated as fixed). Importantly, Assumption 3.1 does not impose that we are sampling from the same underlying distribution across time periods, implying that it is fully compatible with compositional changes (Hong, 2013). This is in contrast to most of the DiD literature. For example, Assumption 1(b) in Sant'Anna and Zhao (2020) explicitly imposes that $(D,X) \perp\!\!\!\perp T$; see also Heckman et al. (1997), and Abadie (2005) for other DiD procedures that rely on this stationarity condition.

As is typical in DiD setups, we are interested in the average treatment effect in time period $t = 1$ among the treated units,

$$ATT = \tau = \mathbb{E}\left[Y_1(1)\,|D=1, T=1\right] - \mathbb{E}\left[Y_1(0)\,|D=1, T=1\right]. \tag{3.2.1}$$

Given that the untreated potential outcome $Y_{i1}(0)$ is never observed for the treated units, we need to impose assumptions to uncover $\mathbb{E}\left[Y_1(0)\,|D=1, T=1\right]$ from the data. We make conditional parallel trends, no-anticipation, and strong overlap assumptions toward this goal. Let $\mathscr{S} \equiv \{0,1\}^2$ and $\mathscr{S}_- \equiv \{(1,0),(0,1),(0,0)\}$.

Assumption 3.2 (Conditional Parallel Trends, No-Anticipation, and Overlap)
For some $\varepsilon > 0$, $(d,t) \in \mathscr{S}_-$, and for all $x \in \mathscr{X}$

$$\begin{aligned}
(i) \quad & \mathbb{E}[Y_1(0)|D=1,T=1,X=x] - \mathbb{E}[Y_0(0)|D=1,T=0,X=x] \\
& \qquad = \mathbb{E}[Y_1(0)|D=0,T=1,X=x] - \mathbb{E}[Y_0(0)|D=0,T=0,X=x]. \\
(ii) \quad & \mathbb{E}[Y_0(0)|D=1,T=0,X=x] = \mathbb{E}[Y_0(1)|D=1,T=0,X=x]. \\
(iii) \quad & \mathbb{P}(D=1,T=1) > \varepsilon \text{ and } \mathbb{P}(D=d,T=t|X=x) \geq \varepsilon.
\end{aligned}$$

Assumption 3.2(i) is the conditional parallel trends assumption (CPT) stating that conditioning on X, the average evolution of the untreated potential outcome is the same among the treated and untreated groups. This assumption allows for covariate-specific trends and does not restrict the trends among different covariate strata. Assumption 3.2(ii) is a no-anticipation assumption (NAA) stating that, on average, treated units do not act on the future treatment prior to its implementation (Abbring and van den Berg, 2003; Malani and Reif, 2015). Assumption 3.2(iii) is an overlap condition that guarantees that there are some treated units in the post-treatment period and that the covariates do not fully determine treatment status. This condition is crucial for guaranteeing nonparametric regular inference procedures (Khan and Tamer, 2010).

3.2.2 Identification and Semiparametric Efficiency Bound

Under Assumptions 3.1 and 3.2, it is straightforward to show that the ATT is nonparametrically identified by the outcome regression estimand[2]

$$\tau = \tau_{or} \equiv \mathbb{E}\left[Y|D=1,T=1\right] - \mathbb{E}\left[m_{1,0}(X) + m_{0,1}(X) - m_{0,0}(X)|D=1,T=1\right], \tag{3.2.2}$$

where $m_{d,t}(x) = E[Y|D=d,T=t,X=x]$. Alternatively, it is also easy to show that one can identify the ATT using an inverse probability weighted estimand

$$\tau = \tau_{ipw} \equiv \mathbb{E}\left[\left(w_{1,1}(D,T) - w_{1,0}(D,T,X) - w_{0,1}(D,T,X) + w_{0,0}(D,T,X)\right)Y\right], \tag{3.2.3}$$

where, for $(d,t) \in \mathscr{S}_{-}$

$$\begin{aligned} w_{1,1}(D,T) &= \frac{DT}{\mathbb{E}[DT]}, \\ w_{d,t}(D,T,X) &= \frac{I_{d,t} \cdot p(1,1,X)}{p(d,t,X)} \Big/ \mathbb{E}\left[\frac{I_{d,t} \cdot p(1,1,X)}{p(d,t,X)}\right], \end{aligned} \tag{3.2.4}$$

$I_{d,t} = \mathbb{1}\{D=d,T=t\}$, and $p(d,t,x) = \mathbb{P}(D=d,T=t|X=x)$ is a so-called generalized propensity score. Notice that the weights in (3.2.4) are of the Hájek (1971)-type. This guarantees that all the weights sum up to one and typically results in more stable finite sample behavior; see, e.g., Millimet and Tchernis (2009); Busso et al. (2014); Sant'Anna and Zhao (2020).

From (3.2.2) and (3.2.3), it is clear that any linear combination of τ_{or} and τ_{ipw} also identifies the ATT under our assumptions. There are also many other potential estimands that make use of nonlinear combinations of the different terms in τ_{or} and τ_{ipw} and identify the ATT. From this simple observation, a natural question that arises is: How can we combine these two strategies to obtain an efficient estimator for the ATT? The next theorem addresses this question through the lens of semiparametric efficiency theory. Specifically, we derive the efficient influence function for the ATT under Assumptions 3.1 and 3.2, as well as its semiparametric efficiency bound. This bound represents the maximum precision achievable in this context under the given assumptions. As so, it provides a benchmark that researchers can use to assess whether any given (regular) semiparametric DiD estimator for the ATT fully exploits the empirical content of Assumptions 3.1 and 3.2.[3] Hereafter, let $\tau(Y,X) = Y - (m_{1,0}(X) + (m_{0,1}(X) - m_{0,0}(X)))$ and $W = (Y,D,X,T)$. We also denote the ATT by τ.

[2] See Lemma 3.2 in Section 3.8.1 for the formalization of these results.

[3] To simplify exposition, we abstract from additional technical discussions related to the conditions to guarantee quadratic mean differentiability and their implications for the precise definition of efficient influence function; see, e.g., Chapter 3 of Bickel et al. (1998) for more details.

Theorem 3.1 (Semiparametric Efficiency Bound) Suppose Assumptions 3.1 and 3.2 hold. Then, the efficient influence function for τ is given by

$$\eta_{\text{eff}}(W) = w_{1,1}(D,T)(\tau(Y,X)-\tau) + \sum_{(d,t)\in\mathscr{S}_-}(-1)^{(d+t)}w_{d,t}(D,T,X)(Y-m_{d,t}(X)), \tag{3.2.5}$$

where the weights are defined in (3.2.4). Furthermore, the semiparametric efficiency bound for the set of all regular estimators of τ is

$$\mathbb{E}[\eta_{\text{eff}}(W)^2] = \frac{1}{\mathbb{E}[DT]^2}\mathbb{E}\left[DT(\tau(Y,X)-\tau)^2 + \sum_{(d,t)\in\mathscr{S}_-}\frac{I_{d,t}\cdot p(1,1,X)^2}{p(d,t,X)^2}(Y-m_{d,t}(X))^2\right].$$

Apart from providing an efficiency benchmark, Theorem 3.1 also provides us a template to construct efficient estimators for τ. That is, given that any influence function has a mean of zero, we can take the expected value of $\eta_{\text{eff}}(W)$ and isolate τ to get the following estimand for the ATT

$$\tau = \tau_{dr} \equiv \mathbb{E}\left[w_{1,1}(D,T)\tau(Y,X) + \sum_{(d,t)\in\mathscr{S}_-}(-1)^{(d+t)}w_{d,t}(D,T,X)(Y-m_{d,t}(X))\right]. \tag{3.2.6}$$

Note that we can rewrite τ_{dr} as the τ_{or} estimand augmented with IPW terms that weight the errors of the regression of Y on X among subgroups defined by $(d,t)\in\mathscr{S}_-$, that is,

$$\tau_{dr} = \tau_{or} + \sum_{(d,t)\in\mathscr{S}_-}(-1)^{(d+t)}\mathbb{E}\left[w_{d,t}(D,T,X)(Y-m_{d,t}(X))\right].$$

Alternatively, one can rewrite τ_{dr} as the τ_{ipw} estimand augmented with re-weighted outcome regression terms.

$$\tau_{dr} = \tau_{ipw} + \sum_{(d,t)\in\mathscr{S}_-}(-1)^{(d+t)}\mathbb{E}\left[(w_{1,1}(D,T)-w_{d,t}(D,T,X))m_{d,t}(X)\right].$$

These alternative representations of the ATT estimand based on the efficient influence function highlight that combining IPW and OR approaches can lead to efficiency gains. In addition, these representations suggest that τ_{dr} possesses the so-called "doubly robust" property, which allows for recovering the ATT, as long as one correctly specifies a model for the generalized propensity score or a model for the outcome regressions. In a nonparametric world, these DR properties can be interpreted as "rate doubly robustness" as shown in Section 3.3.1; see also Smucler et al. (2019).

3.2.3 Bias-Variance Trade-Off With Respect To Stationarity

All the estimands described in Section 3.2.2 account for compositional changes over time, and the τ_{dr} estimand (3.2.6), based on the efficient influence function, inherit efficiency properties under our assumptions. As mentioned in the introduction, most DiD estimators typically assume no compositional changes *a priori*. A natural question then arises: How biased would these estimators be when they erroneously rule out compositional changes?

To tackle this question, we examine the bias of the semiparametrically efficient DiD estimator for the ATT proposed by Sant'Anna and Zhao (2020) that excludes compositional changes. Before diving into this analysis, we need to introduce some additional notation and clarify the assumptions, estimands, and other aspects of Sant'Anna and Zhao (2020)'s approach.

First, Sant'Anna and Zhao (2020) explicitly rules out compositional changes by relying on the following stationarity assumption.

Assumption 3.3 (Stationarity) $(D,X) \perp\!\!\!\perp T$.

Intuitively, Assumption 3.3 enables researchers to pool covariates and treatment variables from both time periods. As a result, under Assumption 3.3, it follows that $\mathbb{E}[D|X,T=1] = \mathbb{E}[D|X] \equiv \tilde{p}(X)$, which also affects the definition of the "relevant" propensity score. Sant'Anna and Zhao (2020) fully exploit these features and show that, under Assumptions 3.1, 3.2, and 3.3, the efficient influence function for the ATT is given by

$$\eta_{sz}(W) = \frac{D}{\mathbb{E}[D]}\Big(\tau(X) - \tau\Big) + \sum_{(d,t)\in\mathscr{S}} (-1)^{(d+t)} w^{sz}_{d,t}(D,T,X)(Y - m_{d,t}(X)), \tag{3.2.7}$$

where $\tau(x) = (m_{1,1}(x) - m_{1,0}(x)) - (m_{0,1}(x) - m_{0,0}(x))$ is the conditional ATT, and for $t = 0,1$,

$$\begin{aligned} w^{sz}_{1,t}(D,T,X) &= \frac{D \cdot \mathbb{1}\{T=t\}}{\mathbb{E}[D\cdot \mathbb{1}\{T=t\}]}, \\ w^{sz}_{0,t}(D,T,X) &= \frac{\tilde{p}(X)(1-D)\cdot \mathbb{1}\{T=t\}}{1-\tilde{p}(X)} \Bigg/ \mathbb{E}\left[\frac{\tilde{p}(X)(1-D)\cdot \mathbb{1}\{T=t\}}{1-\tilde{p}(X)}\right]. \end{aligned} \tag{3.2.8}$$

Based on (3.2.7), Sant'Anna and Zhao (2020) propose the following DR estimand for the ATT:

$$\tau_{sz} \equiv \mathbb{E}\left[\frac{D}{\mathbb{E}[D]}\tau(X) + \sum_{(d,t)\in\mathscr{S}} (-1)^{(d+t)} w^{sz}_{d,t}(D,T,X)(Y - m_{d,t}(X))\right]. \tag{3.2.9}$$

The next proposition shows that τ_{sz} does not recover the ATT when Assumption 3.3 is potentially violated, i.e., under compositional changes. It also precisely quantifies the bias relative to τ_{sz}.

Proposition 3.1 Under Assumptions 3.1 and 3.2, we have that

$$
\begin{aligned}
\tau_{sz} - \tau_{dr} = & \sum_{(d,t)\in\mathscr{S}} (-1)^{(d+t)} \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{DT}{\mathbb{E}[DT]}\right) m_{d,t}(X)\right] \\
& + \sum_{(d,t)\in\mathscr{S}_-} (-1)^{(d+t)} \mathbb{E}\left[\left(w^{sz}_{d,t}(D,T,X) - w_{d,t}(D,T,X)\right)(Y - m_{d,t}(X))\right] \\
= & \mathbb{E}[\tau(X)|D=1] - \mathbb{E}[\tau(X)|D=1, T=1] \\
= & \mathbb{E}[\tau(X)|D=1] - \tau.
\end{aligned}
$$

Proposition 3.1 provides bias decomposition for τ_{sz} when the stationarity assumption is not imposed. The first equality in Proposition 3.1 follows from a direct comparison between our proposed estimand for the ATT and the one proposed by Sant'Anna and Zhao (2020), while the second equality is a consequence of the law of iterated expectations.[4] The third equality is due to the definition of ATT and Assumptions 3.1 and 3.2. These calculations show that Sant'Anna and Zhao (2020)'s DR DiD estimand for the ATT can be biased when Assumption 3.3 is violated. In contrast, our proposed estimand τ_{dr} is fully robust against compositional changes.

Proposition 3.1 also highlights that not all violations of Assumption 3.3 result in biases in ATT when using Sant'Anna and Zhao (2020)'s estimand. Although intuitive and simple, this insight seems to be new in the literature. Based on this observation, one can determine if violations of Assumption 3.3 lead to empirically relevant biases in the ATT by comparing nonparametric estimates based on τ_{sz} with those based on our proposed estimand τ_{dr}. This would detect only the "relevant" violations of Assumption 3.3 that affect the target parameter of interest. That is, it would concentrate power in the directions that one cares about in this context. We discuss this testing procedure in greater detail in Section 3.4.

At this point, one may also wonder what the price one pays for such robustness in terms of semiparametric efficiency. Specifically, how much efficiency one loses by using τ_{dr} when Assumption 3.3 holds but is not fully exploited. The next proposition compares the semiparametric efficiency bound derived in Theorem 3.1 with the one derived by Sant'Anna and Zhao (2020).

Proposition 3.2 (Efficiency Loss under Stationarity) Suppose that Assumptions 3.1, 3.2, and 3.3 hold. Then

$$
\rho_{sz} \equiv \mathbb{E}[\eta_{\text{eff}}(W)^2] - \mathbb{E}[\eta_{sz}(W)^2] = \frac{1-\mathbb{E}[T]}{\mathbb{E}[D]\,\mathbb{E}[T]} \mathbb{V}\text{ar}\left[\tau(X)|D=1\right]. \tag{3.2.10}
$$

[4] Here, we are implicitly considering the case where there are no (global) model misspecifications, which aligns with the fully nonparametric approach we adopt. One can compute a similar bias decomposition when one adopts parametric working models for the nuisance functions, though the notation becomes much more cumbersome.

It is evident from Proposition 3.2 that our proposed estimator is asymptotically less efficient than the one proposed by Sant'Anna and Zhao (2020) when there are no compositional changes over time. The efficiency loss is greater if any of the following three quantities is larger: 1) the population ratio of the pre-treatment period vs. the post-treatment period, 2) the population proportion of the comparison group vs. the treated group, and 3) the expected variability of treatment effect heterogeneity among the treated. In the extreme case where the treatment effect on the treated is homogeneous, our ATT estimator would achieve the same efficiency level as the one that imposes stationarity *a priori*. However, we imagine this case is not very realistic.

Propositions 3.1 and 3.2 characterize a bias-variance trade-off. Although our proposed estimand for the ATT is robust against Assumption 3.3, there is an asymptotic efficiency loss of not exploiting Assumption 3.3 when it does hold. We revisit this trade-off in Section 3.4.

3.3 Estimation and Inference

The results from Section 3.2.2 suggest one can estimate the ATT by building on the efficient influence function derived in Theorem 3.1, as emphasized by (3.2.6). The results from Propositions 3.1 and 3.2 also suggest a testing procedure to assess whether compositional changes translate to biased ATT estimates. However, all the discussions so far has involved estimands that depend on unknown nuisance functions, and we have not yet discussed how one can estimate these to form feasible two-step estimators. This section discusses how to proceed when adopting a fully nonparametric approach, therefore avoiding additional functional form assumptions.

We first present a generic result that emphasizes that estimators based on (3.2.6) have a rate DR property, regardless of how you choose to (nonparametrically) estimate the nuisance functions. Although interesting and useful, this generic result does not help us with practical inference procedures. Towards that end, we discuss how one can concretely estimate the generalized propensity score (PS) and outcome regression (OR) nuisance functions using local polynomials, even in the presence of discrete covariates. We then establish the large sample properties of our DR DiD two-step estimator for the ATT based on local polynomials. We provide a data-driven bandwidth selection method in Subsection 3.3.4. We defer the construction of the Hausman-type test for compositional changes to Section 3.4.

3.3.1 Rate Doubly Robust

Let $\widehat{p}$, and $\widehat{m}_{d,t}$ be generic estimators of p, and $m_{d,t}$, for $(d,t) \in \mathscr{S}_{-}$. Given these first-step estimators, our proposed two-step estimator for the ATT based on (3.2.6) is given by

$$\widehat{\tau}_{dr} = \mathbb{E}_n \left[\widehat{w}_{1,1}(D,T)\widehat{\tau}(Y,X) + \sum_{(d,t)\in\mathscr{S}_{-}} (-1)^{(d+t)} \widehat{w}_{d,t}(D,T,X)(Y - \widehat{m}_{d,t}(X)) \right], \tag{3.3.1}$$

where $\widehat{\tau}(Y,X) = Y - (\widehat{m}_{1,0}(X) + (\widehat{m}_{0,1}(X) - \widehat{m}_{0,0}(X)))$, and, for $(d,t) \in \mathscr{S}_{-}$,

$$\widehat{w}_{1,1}(D,T) = \frac{DT}{\mathbb{E}_n[DT]}, \tag{3.3.2}$$

$$\widehat{w}_{d,t}(D,T,X) = \frac{I_{d,t} \cdot \widehat{p}(1,1,X)}{\widehat{p}(d,t,X)} \Big/ \mathbb{E}_n \left[\frac{I_{d,t} \cdot \widehat{p}(1,1,X)}{\widehat{p}(d,t,X)} \right]. \tag{3.3.3}$$

We impose the following assumptions on the quality of nuisance function estimators. We let $\|f\|_{L_2} \equiv \left(\int f^2 d\mu\right)^{1/2}$ and $\|f\|_\infty \equiv \sup_{x \in \mathscr{X}} |f(x)|$ denote the L_2- and sup-norm of a function f, respectively, and let $\mathbb{G}_n(\cdot)$ denote the empirical process $\sqrt{n}\,(\mathbb{E}_n - \mathbb{E})\,(\cdot)$.

Assumption 3.4 (Estimation of Nuisance Parameters)

1. The estimators $\widehat{p}$ and $\widehat{m}$ are uniformly convergent in the sense that

$$\|\widehat{p}(\cdot,\cdot,\cdot) - p(\cdot,\cdot,\cdot)\|_\infty = o_p(1), \qquad \max_{(d,t) \in \mathscr{S}_{-}} \left\|\widehat{m}_{d,t}(\cdot) - m_{d,t}(\cdot)\right\|_\infty = o_p(1).$$

2. For $(d,t) \in \mathscr{S}_{-}$,

$$\begin{aligned}
&(i) \qquad \mathbb{E}_n[(Y - m_{d,t}(X)) \cdot \left(\widehat{w}_{d,t} - w_{d,t}\right)(W)] = o_p(n^{-1/2}). \\
&(ii) \qquad \mathbb{E}_n[(w_{1,1} - w_{d,t})(W) \cdot \left(\widehat{m}_{d,t} - m_{d,t}\right)(X)] = o_p(n^{-1/2}). \\
&(iii) \qquad \mathbb{G}_n\left\{ I_{d,t} \cdot \left(\frac{\widehat{p}(1,1,X)}{\widehat{p}(d,t,X)} - \frac{p(1,1,X)}{p(d,t,X)} \right) \cdot \left(\widehat{m}_{d,t} - m_{d,t}\right)(X) \right\} = o_p(1). \\
&(iv) \qquad \mathbb{G}_n\left[w_{d,t}(W) \cdot \left(\widehat{m}_{d,t} - m_{d,t}\right)(X) \right] = o_p(1). \\
&(v) \qquad \mathbb{G}_n\left[I_{d,t} \cdot \left(\frac{\widehat{p}(1,1,X)}{\widehat{p}(d,t,X)} - \frac{p(1,1,X)}{p(d,t,X)} \right) \right] = o_p(1).
\end{aligned}$$

One can verify these high-level conditions using empirical process arguments. These typically involve ensuring that the functional space in which the first-stage estimation error resides is not overly complex; see, e.g., Kennedy et al. (2017).

Let $(r_n)_{n \geq 1}$ and $(s_n)_{n \geq 1}$ be positive sequences converging to zero such that

$$\max_{(d,t) \in \mathscr{S}_{-}} \|\widehat{p}(d,t,\cdot) - p(d,t,\cdot)\|_e = O_p(r_n),$$

$$\max_{(d,t) \in \mathscr{S}_{-}} \left\|\widehat{m}_{d,t}(\cdot) - m_{d,t}(\cdot)\right\|_e = O_p(s_n),$$

where $e = L_2$ or ∞.

Lemma 3.1 (Doubly-Robust Error Rate with Generic First Step Estimators) Suppose that $e = \infty$, and that Assumptions 3.1, 3.2, 3.4.1 and 3.4.2 (i, ii) are satisfied. Then,

$$\widehat{\tau}_{dr} - \tau = \frac{1}{n}\sum_{i=1}^{n} \eta_{\text{eff}}(W_i) + O_p\left(r_n s_n\right) + o_p\left(n^{-1/2}\right). \tag{3.3.4}$$

Furthermore, if Assumptions 3.4.2 (iii)–(v) are also fulfilled, the equation (3.3.4) remains valid when $e = L_2$.

The lemma demonstrates that our estimator is doubly robust in terms of its convergence rate. The remaining term is the product of the error rates of the first-stage estimators. Due to the product structure, each estimator typically needs only to converge to its true value at a rate of $o(n^{-1/4})$ for the ATT estimator to converge at the parametric rate. This property also allows for a trade-off between precision in the two nuisance estimators.

In the following subsection, we present lower-level conditions for cases in which the nuisance functions are estimated nonparametrically using “leave-one-out” local polynomial estimators. The ‘leave-one-out’ technique enables us to directly establish the conditions in Assumption 3.4.2 without relying on empirical process theory. This is desirable, as verifying the complexity of the space where local polynomial (logistic) estimators reside is not a trivial task.

3.3.2 Local Polynomial Estimation of Nuisance Functions

We first introduce the estimator for the PS functions. Conditional probability functions are naturally bounded within the unit interval. However, these bounds may not be respected when using linear probability models. As a nonparametric generalization of parametric multinomial logit regression, local multinomial logit regression enforces such bounds by design. Through extensive Monte Carlo simulations, Frölich (2006) demonstrates that the local multinomial logit estimator consistently outperforms local least squares, Klein–Spady, and Nadaraya–Watson estimators. Hence, we prefer this estimator over other nonparametric methods.

Let us assume that there are functions $\{g_{d,t}(\cdot)\}_{(d,t)\in\mathscr{S}_-}$, such that

$$p(d,t,x) = \frac{\exp(g_{d,t}(x))}{1+\sum_{(d',t')\in\mathscr{S}_-}\exp(g_{d',t'}(x))},$$

for $(d,t) \in \mathscr{S}_-$, and $p(1,1,x) = \left(1+\sum_{(d',t')\in\mathscr{S}_-}\exp(g_{d',t'}(x))\right)^{-1}$. That is, we suppose that the generalized PS can be represented by a multinomial logistic transformation of unknown functions $\{g_{d,t}(\cdot)\}_{(d,t)\in\mathscr{S}}$. Instead of imposing specific functional forms on $\{g_{d,t}(\cdot)\}_{(d,t)\in\mathscr{S}_-}$, the local multinomial logit estimator approximates these unknown functions locally using polynomials, which we will describe in detail below.

In line with the conventions of local polynomial estimation, we employ the following notations as shorthand for common operators on vectors,

$$\mathbf{k}=(k_1,...,k_v),\quad |\mathbf{k}|=\sum_{\ell=1}^{v}k_\ell,\quad \mathbf{k}!=\prod_{\ell=1}^{v}k_\ell!,\quad x^{\mathbf{k}}=\prod_{\ell=1}^{v}x_\ell^{k_\ell},$$

$$f^{(\mathbf{k})}(x)=\frac{\partial^{\mathbf{k}}f(x)}{\partial x_1^{k_1}\cdot\partial x_2^{k_2}\cdots\partial x_v^{k_v}},\qquad \sum_{0\le|\mathbf{k}|\le p}f(\mathbf{k})=\sum_{\ell=0}^{p}\sum_{k_1=0}^{\ell}\ldots\sum_{\substack{k_v=0\\k_1+\ldots+k_v=\ell}}^{\ell}f(k_1,...,k_v).$$

Furthermore, we define $n_k=\binom{k+\ell-1}{\ell-1}$ as the number of distinct ℓ-tuples $\mathbf{k}$ with $|\mathbf{k}|=k$. We arrange these n_k ℓ-tuples in a lexicographically-ordered sequence, prioritizing the last position, and denote the mapping from the rank in the ordered sequence to the corresponding ℓ-tuple as $\pi_k(\cdot)$.

Our method accommodates discrete and continuous covariates, so we must differentiate between these variables. We assume that $x=(x_c,x_d)$, where x_c is a υ_c-vector of continuous covariates, and x_d is the subvector of discrete variables. We also distinguish between ordered and unordered discrete variables. That is, $x_d=(x_u,x_o)$, where x_u is a υ_u-vector of unordered covariates and x_o is a υ_o-vector of ordered covariates.

Now, for a generic function, $g:\mathscr{X}\to\mathbb{R}$, and a point, $x^*\in\mathscr{X}$, $g(\cdot)$ can be approximated in a neighborhood of x^* by a p-th order Taylor series with respect to the continuous variables, as

$$g(x)\approx\sum_{0\le|\mathbf{k}|\le p}\frac{1}{\mathbf{k}!}g^{(\mathbf{k})}(x^*)(x_c-x_c^*)^{\mathbf{k}}=\underline{\mathbf{X}}(x_c^*)'\gamma_g(x^*),$$

where $\underline{\mathbf{X}}_p(x_c)=(\underline{\mathbf{X}}^{(0)\prime}(x_c),...,\underline{\mathbf{X}}^{(p)\prime}(x_c))'$ is a $N_p\times 1$ vector that contains the sorted $(X_c-x_c)^{\mathbf{k}}$, with $N_p\equiv\sum_{k=0}^{p}n_k$. The l-th entry of $\underline{\mathbf{X}}^{(k)}(x_c)$, denoted as $\underline{\mathbf{X}}^{(k,l)}(x_c)$, is equal to $(X_c-x_c)^{\pi_k(l)}$. The vector $\gamma_g(x)=(\gamma_g^{(0)'}(x),\ldots,\gamma_g^{(p)'}(x))'$ is defined as the vector of lexicographically-ordered $g^{(\mathbf{k})}(x)/\mathbf{k}!$.

The local approximation is achieved through kernel smoothing. For continuous variables, we let the kernel function be denoted by $K^j(\mathbf{u})$, $j=ps,or$. It is a nonnegative function supported on $[-1,1]^{\upsilon_c}$. Suppose $h>0$ is a generic bandwidth parameter. We denote the scaled kernel function by $K_h(\mathbf{u})=K(\mathbf{u}/h)/h^{\upsilon_c}$. We use the kernel function proposed by Li and Racine (2007) for discrete variables. This kernel function is defined as

$$L_\lambda(x_d,z_d)=\prod_{s=1}^{\upsilon_u}\lambda_u^{\mathbb{1}\{x_{u,s}-z_{u,s}\}}\prod_{s=1}^{\upsilon_o}\lambda_o^{|x_{o,s}-z_{o,s}|},\tag{3.3.5}$$

where $\lambda=(\lambda_u,\lambda_o)\in[0,1]^2$ is a generic smoothing parameter. When $\lambda=0$, the estimator reduces to the frequency estimator.

For the j-th observation of covariates, X_j, our local polynomial (multinomial) logit estimator of γ, denoted by $\widehat{\gamma}$, satisfies

$$\widehat{\gamma}(X_j) \equiv (\widehat{\gamma}_{1,0}(X_j), \widehat{\gamma}_{0,1}(X_j), \widehat{\gamma}_{0,0}(X_j))' = \arg\max_{\gamma \in \mathbb{R}^{3N_p}} \frac{1}{n-1} \sum_{i \neq j}^{n} \ell(W_i, X_j; \gamma) \widetilde{K}_{ps}(X_i; X_j, h, \lambda), \tag{3.3.6}$$

where $\widetilde{K}_{ps}(X_i; X_j, h, \lambda) = K_h^{ps}(X_{c,i} - X_{c,j}) L_\lambda(X_d, X_{d,j})$ and the local likelihood function $\ell(w, x; \gamma)$ is defined as

$$\ell(w, x; \gamma) = \sum_{(d',t') \in \mathscr{S}_-} I_{d,t} \underline{\mathbf{X}}_p(x_c)' \gamma_{d,t} - \log\left(1 + \sum_{(d',t') \in \mathscr{S}_-} \exp\left(\underline{\mathbf{X}}_p(x_c)' \gamma_{d',t'}\right)\right).$$

Note that we have used a "leave-one-out" version of the local regression estimator for the construction of $\widehat{\gamma}$, i.e., $\gamma(X_j)$ are estimated using every observation except the j-th. This technique, standard in the literature (Powell and Stoker, 1996; Powell et al., 1989; Rothe and Firpo, 2019), serves to avoid a "leave-in" bias that is of first-order importance when estimating the ATT.

Let $e_{\ell,k}$ denote an ℓ-dimensional vector in which the k-th element is set to one, while all remaining elements are zero. Then, for a given $\widehat{\gamma}$, the generalized PS can be approximated by[5]

$$\widehat{p}(d, t, x) = \frac{\exp(e'_{N_p,1} \widehat{\gamma}_{d,t}(x))}{1 + \sum_{(d',t') \in \mathscr{S}_-} \exp(e'_{N_p,1} \widehat{\gamma}_{d',t'}(x))}, \tag{3.3.7}$$

for $(d,t) \in \mathscr{S}_-$, and $\widehat{p}(1,1,x) = 1 - \sum_{(d,t) \in \mathscr{S}_-} \widehat{p}(d,t,x)$.

For OR models, we employ leave-one-out q-th order local polynomial least squares estimators. First, the local polynomial regression coefficients are estimated by solving the following equation:

$$\widehat{\beta}_{d,t}(X_j) = \arg\min_{\beta \in \mathbb{R}^{N_p}} \frac{1}{n-1} \sum_{i \neq j}^{n} \left(Y_i - \underline{\mathbf{X}}_{q,i}(X_{c,j})' \beta\right)^2 I_{d,t,i} \widetilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}), \tag{3.3.8}$$

where $\widetilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}) = K_{b_{d,t}}^{or}(X_{c,i} - X_{c,j}) L_{\vartheta_{d,t}}(X_d, X_{d,j})$, and $I_{d,t,i} = \mathbb{1}\{D_i = d, T_i = t\}$. Then, we estimate the OR functions by

$$\widehat{m}_{d,t}(X_j) = e'_{N_q,1} \widehat{\beta}_{d,t}(X_j), \tag{3.3.9}$$

for $(d,t) \in \mathscr{S}_-$.

We analyze the asymptotic behaviors of these local polynomial estimators in Section 3.8.2. We provide results on the uniform convergence rate for the approximation error. In particular, we establish a uniform stochastic expansion for the local multinomial logit regression that is of independent interest.

Remark 3.1 The choice of polynomial order depends on considerations such as computational tractability and the

[5] We abuse notation and denote the local polynomial estimators for the generalized propensity score as $\widehat{p}$ and for the outcome regression as $\widehat{m}$, which are the same as the generic estimators introduced in Section 3.3.1.

trade-off between bias and variance properties. We adhere to the recommendation made by Fan et al. (1995) to employ odd-degree polynomial fits, as they simplify the analysis for the boundary bias when using symmetric kernel functions. We allow varying local polynomial orders for the PS and OR estimators and, in the case of the latter, for distinct treatment groups. This flexibility is desirable as the propensity score and conditional mean functions might display varying degrees of smoothness.

3.3.3 Asymptotic Normality

With $\{\widehat{m}_{d,t}\}_{(d,t)\in\mathscr{S}_-}$ given in (3.3.9), and $\widehat{p}$ defined in (3.3.7), we can construct an estimator for τ_{dr} as shown in (3.3.1). In the following, we derive the large sample properties of the estimator $\widehat{\tau}_{dr}$ by applying Lemma 3.1. To achieve this objective, we begin by presenting a set of regularity assumptions. Henceforth, we use $\mathscr{B}(x,\delta)$ to denote a ball centered at x with radius δ, and $\lambda_{min}(A)$ to represent the smallest eigenvalue of a square matrix A.

Assumption 3.5 (Support, Smoothness, Integrability, Kernel, and Bandwidth conditions)

1. (i) $\mathscr{X} = \mathscr{X}_c \otimes \mathscr{X}_d$, where $\mathscr{X}_c$ is a compact subset of $\mathbb{R}^{\upsilon_c}$ and $\mathscr{X}_d$ is finite; (ii) For all $x_d \in \mathscr{X}_d$, $\mathbb{P}(X_d = x_d) > 0$, and the conditional probability density of X_c, $f_{X_c|X_d}(\cdot|x_d)$, is continuously differentiable and bounded away from zero on $\mathscr{X}_c$; (iii) There are positive constants κ_0 and κ_1 in $(0,1]$ such that for any $x \in \mathscr{X}$ and all $\varepsilon \in (0,\kappa_0]$, there exists a $x' \in \mathscr{X}$ satisfying, $x'_d = x_d$, and
$$\mathscr{B}\left(x',\kappa_1\varepsilon\right) \subset \mathscr{B}(x,\varepsilon) \cap \mathscr{X}.$$

2. For all $x \in \mathscr{X}$, (i) $p(d,t,x)$ is $(p+1)$-times continuously differentiable in x_c, with uniformly bounded derivatives, for $(d,t) \in \mathscr{S}$; (ii) $m_{d,t}(x)$ is $(q+1)$-times continuously differentiable in x_c, with uniformly bounded derivatives, for $(d,t) \in \mathscr{S}_-$.

3. $\mathbb{E}[|Y|^{\zeta}|X,D,T] < \infty$ a.s. for some constant $\zeta > 2$.

4. For $j = ps, or$, (i) $K^j : [-1,1]^{\upsilon_c} \to \mathbb{R}_+$; (ii) $K^j(\cdot)$ satisfies the Lipschitz condition, i.e. $\left|K^j(\mathbf{u}) - K^j(\mathbf{u}')\right| \leq L\|\mathbf{u}-\mathbf{u}'\|$ for some $L > 0$ and any $\mathbf{u},\mathbf{u}' \in \mathbb{R}^d$.

5. (i) $h = o(1)$; (ii) $\log n/\left(nh^{\upsilon_c+2p}\right) = o(1)$ and $\lambda/h^p = o(1)$; (iii) $h^{p+1} = o\left(n^{-1/4}\right)$ and $\log n/(nh^{\upsilon_c}) = o\left(n^{-1/2}\right)$. For $(d,t) \in \mathscr{S}_-$, (iv) $b_{d,t} = o(1)$; (v) $\log n/\left(n^{1-2/\zeta} b_{d,t}^{\upsilon_c}\right) = o(1)$; (vi) $b_{d,t}^{q+1} = o\left(n^{-1/4}\right)$ and $\log n/\left(nb_{d,t}^{\upsilon_c}\right) = o\left(n^{-1/2}\right)$; (vii) $\lambda, \vartheta_{d,t} = o(n^{-1/4})$.

6. With $\mathbf{Q}_j(x_c)$ defined in (3.8.30), $\inf_{x_c\in\mathscr{X}_c} \lambda_{min}\left(\mathbf{Q}_j(x_c)\right) > 0$, for $j = p,q$.

A few remarks on the assumptions are in order. Assumption 3.5.1 indicates that our local polynomial estimator can handle discrete, categorical data. The final part of the condition, proposed by Fan and Guerre (2016), requires that the boundary of $\mathscr{X}$ is sufficiently dense for the first-stage estimators to exhibit good bias and variance properties near the boundary. Assumption 3.5.2 describes the standard smoothness condition for the nuisance functions. Assumption 3.5.3 is a regularity condition that controls the conditional moments of Y. Assumption 3.5.4 collects the regularity conditions on the kernel functions. We note that different kernels can be used for the propensity score and conditional mean models. In practice, the kernel $K(\cdot)$ usually takes a product form, i.e., $K(\mathbf{u}) = \prod_{i=1}^{\upsilon_c} \mathscr{K}(u_i)$, where $\mathscr{K}(\cdot)$ can be selected from several options, such as triangular, biweight, triweight, or Epanechnikov kernels. However, the Gaussian kernel is ruled out due to the restriction on compact support. Assumption 3.5 compiles the rate condition on the bandwidths. Assumptions 3.5.5 (ii) and (v) are imposed to ensure linear expansions of the local polynomial estimators hold uniformly over $\mathscr{X}$. When Y has finite moments of any order, such as when it has bounded support, Assumption 3.5.5 (v) is implied by Assumption 3.5.5 (vi). Assumptions 3.5.5 (iii), (vi), and (vii) specify rate conditions on the bias and stochastic part of the first step estimation error. The usual $o_p(n^{-1/4})$ rate of convergence for the error applies here.

It is important to note that our estimator builds on the efficient influence function and therefore inherits a doubly robust (DR) property. Without such a DR property, it would typically require more stringent rate conditions on the bias part, which can only be satisfied with higher-order kernel functions. See, for example, Newey (1994) and Lee (2018) for detailed discussion. However, this usually results in estimators being more sensitive to tuning parameters, such as bandwidths.

Remark 3.2 Rothe and Firpo (2019) provides a result that can be applied to weaken the rate conditions on the nuisance functions. They present higher-order expansions of semiparametric two-step DR estimators, demonstrating that if the first-step error's bias and the stochastic components are of order $o_p(n^{-1/6})$, and their product is of order $o_p(n^{-1/2})$, the resulting DR estimator achieves root-n consistency. We will not delve into an in-depth discussion on this topic to maintain focus.

Theorem 3.2 (Asymptotic Normality Doubly Robust Estimator) Under Assumptions 3.1, 3.2, and 3.5, we have

$$\sqrt{n}(\widehat{\tau}_{dr} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\text{eff}}(W_i) + o_p(1) \xrightarrow{d} \mathscr{N}(0, \Omega_{dr}), \tag{3.3.10}$$

where $\Omega_{dr} = \mathbb{E}[\eta_{\text{eff}}(W)^2]$.

Theorem 3.2 states that $\widehat{\tau}_{dr}$ is root-n consistent, and asymptotically normal. It also shows that the estimation error of the nuisance functions does not affect the asymptotic distribution of $\widehat{\tau}_{dr}$. Furthermore, the asymptotic variance of $\widehat{\tau}_{dr}$ is equal to the semiparametric efficiency bound.

The theorem can be applied to calculate confidence intervals for the ATT. To achieve this, we need an estimator of the asymptotic variance, Ω_{dr}. One approach to constructing such an estimator is by using empirical analogs of the influence function or through bootstrapping. Here, we focus on the first method, while a weighted bootstrap procedure that accommodates clustered inference is provided in Section 3.8.3.4. Let

$$\widehat{\eta}_{\text{eff}}(W) = \sum_{(d,t)\in\mathscr{S}_-} (-1)^{d+t}\widehat{w}_{d,t}(D,T,X)(Y-\widehat{m}_{d,t}(X)) + \widehat{w}_{1,1}(D,T,X)(\widehat{\tau}(Y,X)-\widehat{\tau}_{dr}), \tag{3.3.11}$$

and $\widehat{\Omega}_{dr} = \mathbb{E}_n[\widehat{\eta}_{\text{eff}}(W)^2]$. Under mild regularity conditions, the consistency of $\widehat{\Omega}_{dr}$ can be established, with its proof included in that of Theorem 3.3 presented in the following section.

3.3.4 Bandwidth Selection

This subsection addresses the practical selection of bandwidth for the first-step local polynomial estimators. It is well-documented that smoothing parameters have a significant impact on balancing the trade-off between bias and variance. Although robustness checks employing multiple bandwidths can be useful, a reliable data-driven selection rule is often preferred. In the following, we outline two cross-validation procedures for choosing these tuning parameters.

Define the following two criterion functions

$$\begin{aligned}
&C_n^{ls}(h,\lambda,\{b_{d,t},\vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-}) \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{(d,t)\in\mathscr{S}}(I_{d,t,i}-\widehat{p}(d,t,X_i))^2 + \sum_{(d,t)\in\mathscr{S}_-} I_{d,t,i}(Y_i-\widehat{m}_{d,t}(X_i))^2\right\}, &(3.3.12)\\
&C_n^{ml}(h,\lambda,\{b_{d,t},\vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-}) \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n}\left\{-\sum_{(d,t)\in\mathscr{S}} I_{d,t,i}\log(\widehat{p}(d,t,X_i)) + \sum_{(d,t)\in\mathscr{S}_-} I_{d,t,i}(Y_i-\widehat{m}_{d,t}(X_i))^2\right\}. &(3.3.13)
\end{aligned}$$

The least-squares criterion, C_n^{ls}, is a standard choice in the kernel estimation literature. It is based on the sum of the least squares distance between the observed and leave-one-out fitted values for both PS and OR estimators, The second criterion, C_n^{ml}, replaces the PS estimator's least squares sum with that of observed likelihood. This idea of using a likelihood-based criterion in local logistic estimation can be traced back to Staniswalis (1989).

The cross-validated bandwidths, $(\widehat{h}^j, \widehat{\lambda}^j, \{\widehat{b}^j_{d,t}, \widehat{\vartheta}^j_{d,t}\}_{(d,t)\in\mathscr{S}})$, minimizes C_n^j for $j = ls, ml$. In Section 3.8.3.2, we investigate the mean integrated squared error (MISE) properties of the first-step estimators and derive the convergence rates of the optimal bandwidths. For local linear estimation (i.e. $p = q = 1$), optimal bandwidths guarantee that the rate conditions in Assumption 3.5.5 are fulfilled if $\upsilon_c < 4$. However, this result does not impose any restrictions on the number of discrete variables.

Remark 3.3 When combined with local multinomial logit estimation, cross-validation can be computationally demanding. This is partly due to the absence of a closed-form solution for local multinomial logit regression, unlike the local least squares regression. Evaluating the criterion function requires solving n minimization problems, which can be time-consuming, particularly for large datasets. To address this issue, we propose a plug-in method for frequency-based local polynomial estimators, detailed in Algorithm 3.8.1 in Section 3.8.3.3. This algorithm leverages analytical expressions for the MISE, circumventing the computational burden of the cross-validation method. We recommend using this procedure when υ_d is small, and the size of the dataset is substantial.

3.4 Testing for Compositional Changes

Propositions 3.1 and 3.2 reveal that our proposed estimator for the ATT is robust against compositional changes; however, it is less efficient than the DR DiD estimator proposed by Sant'Anna and Zhao (2020) when the covariate-stationarity assumption is correctly imposed. This trade-off suggests a nonparametric Hausman (1978)-type test for the absence of compositional changes can be constructed by comparing our proposed estimator with that of Sant'Anna and Zhao (2020). Although Sant'Anna and Zhao (2020) focus on parametric first-step estimators for the nuisance parameters, we modestly extend their analysis by considering nonparametric first-step estimators in this section.

Before detailing the test construction, we define the null and alternative hypotheses, $\mathbf{H}_0$ and $\mathbf{H}_1$, respectively. Let τ_{dr} and τ_{sz} be as defined in (3.2.6) and (3.2.9), respectively. Here, we aim to test

$$\mathbf{H}_0 : \tau_{sz} = \tau_{dr} \qquad \text{against} \qquad \mathbf{H}_1 : \tau_{sz} \neq \tau_{dr}.$$

Under the null, Sant'Anna and Zhao (2020)'s DR DiD estimand is equal to our proposed estimand, while the alternative is the negation of the null hypothesis. Note that we are not interested in directly testing the stationarity assumption, $(D,X) \perp\!\!\!\perp T$, *per se*, but rather testing how this assumption affects the construction of our target parameter of interest, the ATT in period $t = 1$. This allows our test procedure to concentrate power in directions that are arguably more relevant to our context.

To operationalize this testing procedure without invoking additional parametric assumptions, we need a nonparametric estimator for τ_{sz}, which in turn requires nonparametric estimators for the PS $\tilde{p}(\cdot)$ and the OR functions $m_{d,t}(\cdot)$, $(d,t) \in \mathscr{S}$. For the PS, we can use the local polynomial estimators from Section 3.3.2 to construct an estimator for $\tilde{p}(\cdot)$ as

$$\widehat{\tilde{p}}(X) = \widehat{p}(1,1,X) + \widehat{p}(1,0,X),$$

where $\widehat{p}(1,t,X)$ is given by (3.3.7). We can estimate the OR $m_{d,t}(\cdot)$ as in (3.3.9), though here we note that now we need to estimate all four conditional mean functions and not just three as in Section 3.3. Based on these, we can then nonparametrically estimate τ_{sz} by

$$\widehat{\tau}_{sz} \equiv \mathbb{E}_n\left[\frac{D}{\mathbb{E}_n[D]}\widehat{\tau}(X) + \sum_{(d,t)\in\mathscr{S}}(-1)^{(d+t)}\widehat{w}_{d,t}^{sz}(D,T,X)(Y-\widehat{m}_{d,t}(X))\right]. \tag{3.4.1}$$

where $\widehat{\tau}(x) = (\widehat{m}_{1,1}(x) - \widehat{m}_{1,0}(x)) - (\widehat{m}_{0,1}(x) - \widehat{m}_{0,0}(x))$, and, for $t = 0,1$,

$$\begin{aligned}
\widehat{w}_{1,t}^{sz}(D,T,X) &= \frac{D\cdot\mathbb{1}\{T=t\}}{\mathbb{E}_n[D\cdot\mathbb{1}\{T=t\}]},\\
\widehat{w}_{0,t}^{sz}(D,T,X) &= \frac{\widehat{p}(X)(1-D)\cdot\mathbb{1}\{T=t\}}{1-\widehat{p}(X)}\bigg/\mathbb{E}_n\left[\frac{\widehat{p}(X)(1-D)\cdot\mathbb{1}\{T=t\}}{1-\widehat{p}(X)}\right].
\end{aligned}$$

Given this nonparametric estimator for τ_{sz} and our nonparametric estimator for τ_{dr} in (3.3.1), our test statistic is defined as

$$\mathscr{T}_n = n\widehat{V}_n^{-1}(\widehat{\tau}_{dr} - \widehat{\tau}_{sz})^2, \tag{3.4.2}$$

where

$$\widehat{V}_n \equiv \mathbb{E}_n\left[(\widehat{\eta}_{\text{eff}}(W) - \widehat{\eta}_{sz}(W))^2\right],$$

with $\widehat{\eta}_{\text{eff}}(W)$ defined in (3.3.11) and

$$\widehat{\eta}_{sz}(W) \equiv \frac{D}{\mathbb{E}_n[D]}(\widehat{\tau}(X) - \widehat{\tau}_{sz}) + \sum_{(d,t)\in\mathscr{S}}(-1)^{(d+t)}\widehat{w}_{d,t}^{sz}(D,T,X)(Y-\widehat{m}_{d,t}(X)). \tag{3.4.3}$$

$\widehat{V}_n$ is an estimator for the variance of the difference between the two DiD estimators for the ATT. We note that an alternative estimator for this difference under the null could be constructed based solely on the variances of each DiD estimator, i.e., $\tilde{V}_n = \widehat{\Omega}_{dr} - \widehat{\Omega}_{sz}$, with $\widehat{\Omega}_{dr} = \mathbb{E}_n[\widehat{\eta}_{\text{eff}}(W)^2]$ and $\widehat{\Omega}_{sz} = \mathbb{E}_n[\widehat{\eta}_{sz}(W)^2]$. However, such as estimator may lead to a negative variance estimate in finite samples, which is obviously not plausible. Using $\widehat{V}_n$ bypasses this drawback.

In the following theorem, we characterize the asymptotic behavior of this statistic. Let $c_{1-\alpha}^*$ denote the $(1-\alpha)$-th quantile of the chi-squared distribution with one degree of freedom (i.e. χ_1^2).

Theorem 3.3 (Test of Stationarity) Suppose Assumptions 3.1, 3.2, and 3.5 hold. The following additional conditions are satisfied: (i) Assumptions 3.5.2 (ii) and 3.5.5 (iv)–(vii) are fulfilled for $(d,t) = (1,1)$; (ii)$\mathbb{V}\text{ar}[\tau(X)|D=1] > 0$. Then,

(a) under the null space $\mathbf{H}_0$, $\widehat{V}_n \xrightarrow{p} \rho_{sz} > 0$, and

$$\lim_{n\to\infty} \mathbb{P}\left(\mathscr{T}_n \geq c^*_{1-\alpha}\right) = \alpha; \tag{3.4.4}$$

(b) under the alternative space $\mathbf{H}_1$,

$$\lim_{n\to\infty} \mathbb{P}\left(\mathscr{T}_n \geq c^*_{1-\alpha}\right) = 1. \tag{3.4.5}$$

The theorem states that the test controls size and is consistent. Although not discussed in detail here, it is easy to show that our test also has power against sequences of Pitman-type local alternatives that converge to the null at the parametric rate.

Remark 3.4 It is crucial to recognize that our test should be viewed as a "model validation" instead of a "model selection" procedure. For researchers concerned about the validity of Assumption 3.3, it may be tempting to perform a two-stage test. In the first stage, a Hausman specification test is used to "pretest" for the presence of compositional changes, and then, in the second stage, the usual t-test is conducted based on either $\widehat{\tau}_{dr}$ or $\widehat{\tau}_{sz}$, depending on the outcome of the Hausman-test. However, as demonstrated by Guggenberger (2010a), Guggenberger (2010b) and Roth (2022), such a model-selection procedure can lead to substantial size distortions when using standard inference methods.

3.5 Monte Carlo Simulation Study

In this section, we examine the finite sample properties of our proposed estimators and testing procedure. We conduct two Monte Carlo experiments in this section. In the first experiment, there are compositional changes over time, so Assumption 3.3 is violated. In contrast, the second experiment adheres to this assumption as the joint distribution of covariates and treatment is independent of treatment timing. For each design, we compare our nonparametric DR DiD estimator $\widehat{\tau}_{dr}$ defined in (3.3.1), which is robust against compositional changes and semiparametrically efficient, with the nonparametric extension of Sant'Anna and Zhao (2020)'s estimator $\widehat{\tau}_{sz}$ defined in (3.4.1), which assumes no compositional change, and with the estimates of the regression coefficients, τ_{fe}, associated with two-way fixed effect (TWFE) regression specifications of the type

$$Y = \alpha_1 + \alpha_2 T + \alpha_3 D + \tau_{fe}(T \cdot D) + \theta' X + \varepsilon.$$

We consider two TWFE specifications: 1) a linear specification, where all the covariates X enter linearly, and 2) a saturated specification, where, in addition to the linear terms, quadratic terms of the continuous covariates and all the interactive terms of the covariates are also included. We include the TWFE specifications in our comparison set as they are prominent in empirical work.

We employ local linear ($p, q = 1$) kernel estimators for both the PS and OR functions. As described in Section 3.3.2, the PS is estimated using the local likelihood method with the (multinomial) logistic link function, whereas the OR is estimated using the local least squares estimator. We use the second-order Epanechnikov kernel for the continuous covariates, and for the discrete variables, we use the kernel given in (3.3.5). Bandwidth selections are based on the log-likelihood and least squares distance criteria discussed in Section 3.3.4.

Our experiments involve a sample size of $n = 1000$, with each design undergoing $5,000$ Monte Carlo replications. We evaluate the DiD estimators for the ATT using various metrics: average bias, median bias, root mean square error (RMSE), empirical 95% coverage probability, the average length of a 95% confidence interval, and the average of the plug-in estimator for the asymptotic variance. Confidence intervals are calculated using a normal approximation, with asymptotic variances estimated by their sample analogues. We also compute the semiparametric efficiency bound for each design to gauge the potential loss of efficiency/accuracy associated with using inefficient DiD estimators for the ATT. Lastly, we perform a Hausman-type test as described in Section 3.4 under each design and report the empirical rejection rates.

3.5.1 Simulation 1: Non-Stationary Covariate Distribution

We first consider a scenario in which the stationarity condition is not satisfied. Let $\mathbf{X} = (X_1, X_2, ..., X_6)$, where X_1 and X_2 are drawn from Uniform$[-1, 1]$, X_3 and X_4 are binary variables, following Bernoulli(0.5), and the remaining two, X_5 and X_6, are distributed as Binomial$(3, 0.5)$. The six variables are mutually independent.

Define

$$\begin{aligned}
f_{1,0}^{ps}(X) =& 0.4\sum_{s=1}^{2}(X_s - X_s^2) + 0.2\sum_{k=3}^{6} X_k + 0.1\left(\sum_{j\in\{3,5\}}(-1)^{j+1}X_jX_{j+1}\right.\\
&\left. + \sum_{l=1}^{2}\sum_{l'=3}^{6}(-1)^{l+1}X_lX_{l'} + \sum_{\ell=3}^{4}\sum_{\ell'=5}^{6}(-1)^{\ell+\ell'}X_\ell X_{\ell'}\right),\\
f_{0,1}^{ps}(X) =& 0.4(2X_1 + X_2 + X_1^2 - X_2^2 + X_1X_2)\\
&+ 0.2\sum_{k=3}^{6}(-1)^{k+1}X_k + 0.1\left(\sum_{l=3}^{6} X_2X_l + \sum_{\ell=3}^{4} X_\ell X_6\right),\\
f_{0,0}^{ps}(X) =& 0.4(X_1 + 2X_2 - X_1^2 + X_2^2 - X_1X_2)\\
&+ 0.2\sum_{k=3}^{6}(-1)^{k}X_k + 0.1\left(\sum_{l=3}^{6} X_1X_l + \sum_{\ell=3}^{4} X_\ell X_5\right),
\end{aligned}$$

and for the OR models,

$$
\begin{aligned}
f^{or}_{base}(X) &= f^{or}_{het}(X) = 27.4X_1 + 27.4X_2 + 13.7X_1^2 + 13.7X_2^2 + 13.7X_1X_2, \\
f^{or}_{att}(X) &= 27.4X_1 + 13.7X_2 + 6.85\sum_{k=3}^{6} X_k - 15.
\end{aligned}
$$

We consider the following data generating process

$$
p^{s1}(d,t,X) = \begin{cases} \dfrac{\exp(f^{ps}_{d,t}(X))}{1+\sum_{(d,t)\in\mathscr{S}_-}\exp(f^{ps}_{d,t}(X))}, & \text{if } (d,t)\in\mathscr{S}_- \\ \dfrac{1}{1+\sum_{(d,t)\in\mathscr{S}_-}\exp(f^{ps}_{d,t}(X))}, & \text{if } (d,t)=(1,1). \end{cases}
$$

Let $U \sim \text{Uniform}\,[0,\, 1]$. The treatment groups are assigned as follows

$$
(D,T) = \begin{cases} (1,0), & \text{if } U \le p^{s1}(1,0,X), \\ (0,1), & \text{if } p^{s1}(1,0,X) < U \le p^{s1}(1,0,X)+p^{s1}(0,1,X), \\ (0,0), & \text{if } p^{s1}(1,0,X)+p^{s1}(0,1,X) < U \le 1-p^{s1}(1,1,X), \\ (1,1), & \text{if } 1-p^{s1}(1,1,X) < U. \end{cases}
$$

Next, building on Kang and Schafer (2007), we consider the following potential outcomes

$$
Y_0(j) = 210 + f^{or}_{base}(X) + \varepsilon_{het} + \varepsilon_{j,0}, \text{ for } j = 0,1, \tag{3.5.1}
$$

$$
Y_1(0) = 210 + 2f^{or}_{base}(X) + \varepsilon_{het} + \varepsilon_{0,1}, \tag{3.5.2}
$$

$$
Y_1(1) = 210 + 2f^{or}_{base}(X) + f^{or}_{att}(X) + \varepsilon_{het} + \varepsilon_{1,1}, \tag{3.5.3}
$$

where $\varepsilon_{het} \sim N(D\cdot f^{or}_{het}, 1)$ and $\varepsilon_{d,t}$, $(d,t)\in\mathscr{S}$ are independent standard normal random variables.

Under this design, the covariate distribution does not exhibit time variation. However, the PS function is different in the two cross-sections. The mean absolute difference between $p^{s1}(1,1,X)$ and $p^{s1}(1,0,X)$, as well as between $p^{s1}(0,1,X)$ and $p^{s1}(0,0,X)$, are both approximately 0.125, with the maximum difference reaching up to 0.63. Consequently, we expect all of the estimators except for $\widehat{\tau}_{dr}$ will produce biased results. In addition, the stationarity test is likely to reject the null hypothesis with high probability. The results in Table 3.1 support these claims.

First, results in Table 3.1 suggest that both $\widehat{\tau}_{fe}$ and $\widehat{\tau}_{sz}$ are severely biased under this DGP, while $\widehat{\tau}_{dr}$ exhibits negligible bias on average. Moreover, among the three sets of estimators considered, only our proposed estimator attains the correct coverage rate. This result is robust to the bandwidth selection method. Notably, the performance of

Table 3.1: Monte Carlo results under compositional changes. Sample size: $n = 1,000$.

True value of ATT: 4.31. Semiparametric Efficiency Bound: 1753.6							
	Two-way Fixed Effect Estimators						
	Spec.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\widehat{\tau}_{fe}$	Linear	-10.437	-10.445	10.933	10425.033	0.121	12.633
$\widehat{\tau}_{fe}$	Saturated	-11.176	-11.206	11.579	8797.289	0.045	11.612
	Nonparametric Doubly Robust DiD Estimators for the ATT						
	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\widehat{\tau}_{dr}$	ML	-0.009	-0.010	1.374	1838.495	0.949	5.304
$\widehat{\tau}_{dr}$	LS	-0.013	-0.010	1.379	1848.848	0.949	5.314
$\widehat{\tau}_{sz}$	ML	4.427	4.436	4.543	983.436	0.009	3.884
$\widehat{\tau}_{sz}$	LS	4.427	4.435	4.543	983.746	0.009	3.884
	Hausman-type test						
	CV Crit.	Avg. Test Stats.	Emp. Pow. (0.10)	Emp. Pow. (0.05)	Emp. Pow. (0.01)		
	ML	21.250	0.998	0.996	0.978		
	LS	21.199	0.998	0.995	0.976		

Notes: Simulations based on 5,000 Monte Carlo experiments. $\widehat{\tau}_{fe}$ the TWFE regression estimator, $\widehat{\tau}_{dr}$ is our proposed nonparametric DR DiD estimator (3.3.1), and $\widehat{\tau}_{sz}$ is the nonparametric DR DiD estimator (3.4.1) based on Sant'Anna and Zhao (2020). For TWFE regression, we use a linear specification, "Linear", and a saturated specification, "Saturated". For DR DiD estimators, the PS and the OR models are estimated using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the PS function is selected with the log-likelihood criterion, "ML", and the least squares criterion, "LS", respectively. Lastly, "Spec.", "CV Crit.", "Avg. Bias", "Med. Bias", "RMSE", "Asy. Var.", "Cover.", and "CIL", stand for the specification, cross-validation criterion, average simulated bias, median simulated bias, simulated root-mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. The Hausman-type test statistic is calculated based on (3.4.2). "Avg. Test Stats.", and "Emp. Pow. (α)" stand for the average test statistic, and empirical power of the test with a nominal size α, respectively. See the main text for further details.

the TWFE does not improve with a fully-saturated specification, indicating that incorporating nonlinear terms into a TWFE regression does not generally help in identifying heterogeneous treatment effects. In terms of efficiency, it is worth noting that the asymptotic variance of $\widehat{\tau}_{dr}$ is close to the semiparametric efficiency bound, which corroborates the findings of Theorem 3.2. Regarding the testing performance, our Hausman-type test can effectively distinguish between the two nonparametric DiD estimators with a high degree of certainty, which is in line with our theoretical finding.

3.5.2 Simulation 2: Stationary Covariate Distribution

We now slightly adjust the first design by taking the average of propensity scores over time while keeping all other aspects of the DGP constant. Specifically, we define

$$p^{s2}(d,t,X) = \mathbb{P}^{s1}(T=t)\,(p^{s1}(d,1,X) + p^{s1}(d,0,X)),$$

where $\mathbb{P}^{s1}(T=t) = \mathrm{E}[p^{s1}(1,t,X) + p^{s1}(0,t,X)]$. The treatment groups are then assigned based on the realization of a standard uniform random variable on the unit interval partitioned by $\{p^{s2}(d,t,X)\}_{(d,t)\in\mathscr{S}}$. Furthermore, the potential outcomes are determined by (3.5.1)–(3.5.3). Unlike the first DGP, both the covariate distribution and the propensity score function are stationary in this case. As a result, we anticipate that both $\widehat{\tau}_{dr}$ and $\widehat{\tau}_{sz}$ will be consistent for the true ATT. Furthermore, the empirical rejection rate of the Hausman-type test is expected to converge to the nominal sizes. The Monte Carlo results under this DGP are summarized in Table 3.2.

Table 3.2: Monte Carlo results under no compositional changes. Sample size: $n = 1,000$.

	True value of ATT: 9.13. Semiparametric Efficiency Bound: 796.8						
	Two-way Fixed Effect Estimators						
	Spec.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\widehat{\tau}_{fe}$	Linear	-10.649	-10.672	11.106	9907.607	0.087	12.325
$\widehat{\tau}_{fe}$	Saturated	-10.563	-10.617	10.946	7924.684	0.048	11.026
	Nonparametric Doubly Robust DiD Estimators for the ATT						
	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\widehat{\tau}_{dr}$	ML	-0.007	-0.020	1.323	1721.037	0.946	5.133
$\widehat{\tau}_{dr}$	LS	-0.010	-0.027	1.328	1732.416	0.946	5.139
$\widehat{\tau}_{sz}$	ML	-0.015	-0.024	0.958	926.689	0.953	3.771
$\widehat{\tau}_{sz}$	LS	-0.016	-0.024	0.958	926.821	0.953	3.771
	Hausman-type test						
	CV Crit.	Avg. Test Stats.	Emp. Size (0.10)	Emp. Size (0.05)	Emp. Size (0.01)		
	ML	1.045	0.108	0.055	0.009		
	LS	1.045	0.107	0.056	0.009		

Notes: Simulations based on 5,000 Monte Carlo experiments. $\widehat{\tau}_{fe}$ the TWFE regression estimator, $\widehat{\tau}_{dr}$ is our proposed nonparametric DR DiD estimator (3.3.1), and $\widehat{\tau}_{sz}$ is the nonparametric DR DiD estimator (3.4.1) based on Sant'Anna and Zhao (2020). For TWFE regression, we use a linear specification, "Linear", and a saturated specification, "Saturated". For DR DiD estimators, the PS and the OR models are estimated using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the PS function is selected with the log-likelihood criterion, "ML", and the least squares criterion, "LS", respectively. Lastly, "Spec.", "CV Crit.", "Avg. Bias", "Med. Bias", "RMSE", "Asy. Var.", "Cover.", and "CIL", stand for the specification, cross-validation criterion, average simulated bias, median simulated bias, simulated root-mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. The Hausman-type test statistic is calculated based on (3.4.2). "Avg. Test Stats.", and "Emp. Size (α)" stand for the average test statistic, and empirical size of the test with a nominal size α, respectively. See the main text for further details.

In contrast to the results presented in Table 3.1, both $\widehat{\tau}_{dr}$ and $\widehat{\tau}_{sz}$ exhibit minimal bias, and their confidence intervals achieve nominal coverage. Their performance is consistently good across different bandwidth selection methods. The TWFE estimators, however, continue to show substantial bias and achieve nearly negligible coverage, despite having much wider confidence intervals compared to the DR DiD estimators. This occurs because the true treatment effects are heterogeneous, but TWFE specifications do not account for that (i.e., the models are misspecified). In terms of efficiency, the asymptotic variance of $\widehat{\tau}_{sz}$ is reasonably close to the semiparametric efficiency bound. The asymptotic variance of $\widehat{\tau}_{dr}$ is, on average, 2.2 times larger than the semiparametric efficiency bound (that imposes no-compositional changes), which is still significantly lower than that of the TWFE estimators. Given that Assumption 3.3 holds for this DGP, the null hypothesis $\mathbf{H}_0$ is true. The empirical rejection frequency of our Hausman-type test is nearly identical to its nominal value, highlighting the desirable properties of this testing procedure.

3.6 Empirical Illustration: the Effect of Tariff Reduction on Corruption

In this section, we revisit a study from Sequeira (2016) on the effect of import tariff liberalization on corruption patterns. Prior to the phaseout of high tariffs between South Africa and Mozambique, bribery payment was pervasive, often used to dodge tariff taxes. According to Sequeira and Djankov (2014), bribery payments can be found in approximately 80% of all shipment records in a random sample of tracked shipments before a tariff rate reduction in 2008.

This tariff change is the result of a long-standing trade agreement between South Africa and Mozambique. The agreement, the Southern African Development Community Trade Protocol, was signed in 1996. The protocol established a timeline for import tariff reductions between 2001 and 2015. The most significant reduction occurred in 2008, with the average nominal rate decreasing by 5%. The effect of such a tariff liberalization scheme is considerable, as both the likelihood and the amount of bribe payments experienced a significant decline following the phaseout.

To investigate the causal relationship between tariff rate reduction and changes in bribery patterns, Sequeira (2016) leverages a quasi-experimental variation induced by trade protocol: Not all products were subject to the change in tariff rate during the analysis period, enabling products unaffected by the tariff changes to serve as a control group. It is thus possible to utilize the DiD design to analyze how tariff rate changes affect bribe patterns along trade routes.

Sequeira (2016) collects data on the bribe payment along the trade routes between the two countries from 2007 to 2013. This data set has a repeated cross-section structure. Sequeira (2016) mainly considers the following two TWFE regressions:

$$\begin{aligned}
\text{(Linear)} \quad & y_{it} = \gamma_1 TCCi \times Post + \mu Post + \gamma_2 TCCi + \beta_2 BT_i + \Gamma_i + p_i + w_t + \delta_i + \varepsilon_{it}, \\
\text{(Interactive)} \quad & y_{it} = \gamma_1 TCCi \times Post + \mu Post + \gamma_2 TCCi + \beta_2 BT_i + \Gamma_i + \Gamma_i \times Post \\
& \quad + p_i + w_t + \delta_i + \varepsilon_{it},
\end{aligned}$$

where TCC_i and BT_i denote Tariff Change Category and Baseline Tariff, respectively, and y_{it} is one of the measurements of bribery payments for shipment i in period t. TCC is the treatment indicator, which takes value one if the product shipped experienced a tariff reduction in 2008, and zero otherwise. The post-treatment period indicator, $Post$, is equal to one for the years following 2008. BT refers to the tariff rates before 2008. A vector of covariates, Γ, industry, year, and clearing agent fixed effects, p, ω, δ, are also included in the regressions. The interactive specification differs from the linear one by an interaction of $Post$ and the covariates, Γ.

Sequeira (2016) focuses on interpreting γ_1 in both specifications as an estimate of the ATT. However, this interpretation might not be valid when treatment effects are heterogeneous (Meyer, 1995; Abadie, 2005). Our proposed

DR DiD estimator, τ_{dr}, and the one based on Sant'Anna and Zhao (2020), $\widehat{\tau}_{sz}$, could be better suited for the task of identifying and consistently estimating the ATT in the present context. In what follows, we estimate the ATT using our proposed DR DiD estimator and compare the results to those obtained by Sequeira (2016).

To achieve this, we first estimate the PS and OR functions based on local linear logistic regression and local linear OLS, respectively. Following Sequeira (2016), we consider four different outcome measures: a binary variable denoting if a bribe is paid, the logarithmic form, $log(x+1)$, of the amount of bribe payment, the logarithmic form of the amount of bribe paid as a share of the value of the shipment, and as a share of the weight of the shipment, respectively. Across all four specifications, we include the following common covariates: baseline tariff rate, dummy variables for whether the shipper is a large firm, whether the product is perishable, differentiated, an agricultural good, whether the shipments are pre-inspected at origin, monitored, and originates from South Africa. Additionally, we include the day of arrival during the week and the terminal where the cargo was cleared. Our procedures allow for these covariate-specific trends, so the CPT Assumption 3.2(i) holds only after accounting for these observed characteristics. To avoid weak-overlap problems, we truncate PS estimates below 0.01.

Table 3.3 summarizes our results. For each estimator, we report both the unclustered standard errors based on asymptotic approximation (in parentheses) and the cluster-robust standard errors based on the bootstrap procedure in Algorithm 3.8.2 (in brackets), where we cluster at the four-digit HS code level as in Sequeira (2016). Likewise, we conduct two sets of Hausman-type tests – one using unclustered influence functions based on (3.4.2) and the other that accounts for clustering using a bootstrap procedure given in Algorithm 3.8.3.

We first observe that the point estimates are negative for all measures of bribery payment, consistent with the findings of Sequeira (2016). The results based on the two DR DiD methods are generally close to the TWFE estimates with the interactive specification. For instance, we find that a tariff reduction reduces the probability of paying a bribe by 28 to 43 percentage points, depending on the specific estimator used. The result is statistically and economically significant at the usual levels. Tariff reduction also seems to lead to a decrease in bribery.[6] The magnitude of the causal effects based on the weighted results, on the other hand, is more mixed.[7] Results based on the TWFE and DR DiD with no-compositional changes estimators suggest that tariff reduction leads to a statistically significant reduction in the average log of the ratio between bribery payment and shipment values of similar magnitude, while our proposed DR DiD estimator that is robust to compositional changes suggests a twice-as-large effect. When the log of the ratio between bribery payment and tonnage is considered, both nonparametric DR DiD estimators report large yet insignificant (at 95% level) ATT estimates. The results of the Hausman-type test displayed at the bottom of Table

[6] Some of local linear OR estimates were a bit sensitive to bandwidth choice. This is arguably due to the limited number of observations within certain strata. To improve the stability of cross-validation, we impose a common bandwidth across all four treatment groups for each type of covariates.

[7] We avoid attaching a precise interpretation of these log transformations due to the issues raised by Chen and Roth (2023).

Table 3.3: Difference-in-differences estimation results for Sequeira (2016)

Estimator/Outcome	Prob(bribe)	Log(1 + bribe)	Log(1 + bribe/shpt.val.)	Log(1 + bribe/shpt.tonn.)
TWFE - Linear Spec.	-0.429	-3.748	-0.011	-1.914
	(0.083)	(0.724)	(0.003)	(0.341)
	[0.131]	[1.064]	[0.003]	[0.496]
TWFE - Interactive Spec.	-0.296	-2.928	-0.010	-1.597
	(0.082)	(0.746)	(0.004)	(0.402)
	[0.124]	[0.917]	[0.004]	[0.457]
DR DiD $\widehat{\tau}_{sz}$	-0.275	-2.542	-0.014	-0.918
(no-compositional changes)	(0.067)	(0.636)	(0.005)	(0.451)
	[0.096]	[0.773]	[0.006]	[0.492]
DR DiD $\widehat{\tau}_{dr}$	-0.307	-2.888	-0.027	-1.131
(robust to compositional changes)	(0.084)	(0.798)	(0.010)	(0.602)
	[0.109]	[0.915]	[0.014]	[0.635]
Hausman-tests for no-compositional changes				
Unclustered p-value	0.270	0.199	0.084	0.601
Clustered p-value	0.338	0.238	0.175	0.643

Notes: Same data used by Sequeira (2016). The results represent the estimated ATT of tariff rate reduction on bribery payment behavior. Columns 2 through 5 denote estimates for dependent variables representing whether a bribe is paid, the logarithmic form, $log(x+1)$, of the amount of bribe paid, the logarithmic form of the amount of bribe paid as a share of the value of the shipment, and as a share of the weight of the shipment, respectively. We compare four different DiD estimators for the ATT: 1. the two-way fixed effect estimator based on specifications in Column (1) of Tables 8-11 in Sequeira (2016); 2. the two-way fixed effect estimator based on Column (2) from Tables 8-11 in Sequeira (2016); 3. DR DiD estimator based on (3.4.1), and 4. DR DiD estimator based on (3.3.1). The same set of covariates is used for the last two estimators. See the main text for further details on the covariates. Continuous variables are re-scaled between 0 and 1, and then added in with binary variables. For DR DiD estimators, the PS and the OR models are estimated nonparametrically, using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the local linear logistic regression is selected with the log-likelihood criterion. Numbers in the parentheses are unclustered standard errors based on asymptotic approximation. Numbers in brackets refer to standard errors clustered at the level of four-digit HS code. Cluster-robust standard errors are calculated following Algorithm 3.8.2 with 9999 bootstrap draws. Hausman-tests are calculated based on (3.4.2). The clustered p-values are calculated following the bootstrap procedure in Algorithm 3.8.3 with 9999 bootstrap draws. To avoid weak-overlap problems, we truncate PS estimates below 0.01.

3.3 suggest that we lack statistical evidence against the assumption of no-compositional changes, especially when one clusters the standard errors.

In sum, our results support the conclusion of Sequeira (2016) that tariff liberalization decreases corruption. Our DR DiD estimates suggest the size of the effects is approximately the same as that of the original paper, indicating that ruling out treatment effect heterogeneity and compositional changes are not of primary concern in this particular application.

3.7 Concluding Remarks

In this paper, we developed a doubly robust estimator for the ATT within the difference-in-differences framework, allowing for time-varying covariates. We established large sample properties for the proposed estimator when the nuisance functions are estimated nonparametrically. In particular, we derived novel results on the uniform linear expansion of the local multinomial logit estimator with mixed data. We provided extensive discussions comparing our proposed DR estimator with those developed by Sant'Anna and Zhao (2020). Additionally, we proposed a Hausman-type test for assessing the validity of the ATT estimators under consideration. We assessed the finite sample performance of our estimation methods and tests using Monte Carlo simulations. All the finite sample findings are consistent with the asymptotic results. Furthermore, we demonstrated the practical utility of our approach with an empirical application concerning the impact of tariff liberalization on corruption.

An intriguing extension of our work is to the case when the number of time periods is greater than two and when the treatment adoption is staggered, as discussed in Callaway and Sant'Anna (2021). In such contexts, they demonstrate that a family of group-time average treatment effects and their aggregates can be identified under a general no-compositional-change assumption. Allowing for compositional changes in that setup appears promising, particularly since multiple time periods suggest that a no-compositional change assumption may be even more restrictive than in the simple two-period case.

3.8 Supplementary Appendix

This supplemental appendix contains auxiliary lemmas, proofs of the main theorems, and additional results presented in the main text.

Notation: Hereafter, we use the abbreviations CLT, CMT, LIE, and LLN to represent the central limit theorem, continuous mapping theorem, law of iterated expectations, and law of large numbers, respectively. Let $f_X(x) = f_{X_c|X_d}(x_c|x_d) \cdot \mathbb{P}(X_d = x_d)$, $\mathbb{N}_n = \{1,2,...,n\}$, and $\iota(d,t) = \mathbb{1}\{d=1,t=0\} + 2\cdot\mathbb{1}\{d=0,t=1\} + 3\cdot\mathbb{1}\{d=0,t=0\}$. The notation $a_n \lesssim b_n$ implies that $a_n \leq cb_n$ for some positive constant c when n is sufficiently large. The symbol $a_n \sim b_n$ denotes that $a_n/b_n \to 1$ as $n \to \infty$. We define $f \in L_2(\mathscr{U})$ to indicate that $\int_{\mathscr{U}} f^2 d\mu$ is finite, and let the L_2- and sup-norm of f to denote $\|f\|_{L_2}$ and $\|f\|_\infty$, respectively. Denote the ATT by τ, i.e.,

$$ATT = \tau = \mathbb{E}\left[Y_1(1)\,|D=1,T=1\right] - \mathbb{E}\left[Y_1(0)\,|D=1,T=1\right].$$

3.8.1 Proofs for Results from Main Text

Let

$$\tau_{or} = \mathbb{E}\left[Y|D=1,T=1\right] - \mathbb{E}\left[m_{1,0}(X) + m_{0,1}(X) - m_{0,0}(X)|D=1,T=1\right],$$

where $m_{d,t}(x) = E[Y|D=d,T=t,X=x]$, and

$$\tau_{ipw} = \mathbb{E}\left[\left(w_{1,1}(D,T) - w_{1,0}(D,T,X) - w_{0,1}(D,T,X) + w_{0,0}(D,T,X)\right)Y\right],$$

where, for $(d,t) \in \mathscr{S}_-$,

$$w_{1,1}(D,T) = \frac{DT}{\mathbb{E}[DT]},$$

$$w_{d,t}(D,T,X) = \frac{\mathbb{1}\{D=d,T=t\}p(1,1,X)}{p(d,t,X)} \Big/ \mathbb{E}\left[\frac{\mathbb{1}\{D=d,T=t\}p(1,1,X)}{p(d,t,X)}\right],$$

and $p(d,t,x) = \mathbb{P}(D=d,T=t|X=x)$ is a so-called generalized propensity score.

Lemma 3.2 Under Assumptions 3.1 and 3.2, it follows that $\tau_{or} = \tau_{ipw} = \tau$.

Proof of Lemma 3.2:

Outcome regression estimand: Using $m_{d,t}(\cdot) = \mathbb{E}[Y_t(d)|D=d,T=t,X=\cdot]$, $(d,t) \in \mathscr{S}_-$, we get

$$\begin{aligned}
\tau_{or} =& \mathbb{E}[Y_1(1)|D=1,T=1] - \mathbb{E}[\mathbb{E}[Y_0(1)|D=1,T=0,X=x]|D=1,T=1] \\
&+ \sum_{t\in\{0,1\}} (-1)^t\, \mathbb{E}[\mathbb{E}[Y_t(0)|D=0,T=t,X=x]|D=1,T=1] \\
=& \mathbb{E}[Y_1(1)|D=1,T=1] - \mathbb{E}[\mathbb{E}[Y_0(0)|D=1,T=0,X=x]|D=1,T=1] \\
&+ \sum_{t\in\{0,1\}} (-1)^t\, \mathbb{E}[\mathbb{E}[Y_t(0)|D=0,T=t,X=x]|D=1,T=1] \\
=& \mathbb{E}[Y_1(1) - Y_1(0)|D=1,T=1] = \tau,
\end{aligned}$$

where the second equality follows from Assumptions 3.2 (ii) and the third holds under Assumptions 3.2 (i).

Propensity score estimand: Let $p(1,1) = \mathbb{P}(D=1,T=1)$. Under the overlapping conditions in Assumption 3.2(iii), $w_{d,t}(d',t',x)$ are well defined for $(d,t) \in \mathscr{S}_-$, $(d',t') \in \{0,1\}^2$, and $x \in \mathscr{X}$ almost everywhere. Additionally,

$$\begin{aligned}
\mathbb{E}[w_{d,t}(D,T,X)Y] &= \mathbb{E}\left[\frac{p(1,1,X)YI_{d,t}}{p(d,t,X)} \Big/ \mathbb{E}\left[\frac{\mathbb{1}\{D=d,T=t\}p(1,1,X)}{p(d,t,X)}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}[Y|D=d,T=t,X] \cdot \frac{I_{d,t}}{p(d,t,X)}\middle| X\right] \cdot \frac{p(1,1,X)}{p(1,1)}\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y|D=d,T=t,X] \cdot \frac{p(1,1,X)}{p(1,1)}\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y|D=d,T=t,X]|D=1,T=1\right] \\
&= \mathbb{E}\left[m_{d,t}(X)|D=1,T=1\right],
\end{aligned}$$

for $(d,t) \in \mathscr{S}_-$. The second line follows by the LIE, the third equality is by the definition of propensity scores, and the next to last line is by Bayes' Law. Next, from $\mathbb{E}[w_{1,1}(D,T)Y] = \mathbb{E}[Y|D=1,T=1]$ and the same arguments for the OR estimand, we conclude that $\tau_{ipw} = \tau$. ■

Proof of Theorem 3.1:

We follow the steps in Hahn (1998) for the derivation of the efficient influence function. Let $f(y|d,t,x) = f(y|D=$

$d, T = t, X = x)$.

Step 1: characterize the tangent space of the statistical model. The observed likelihood is given as

$$\begin{aligned} f(y,d,t,x) =& f(y|1,1,x)^{dt} f(y|1,0,x)^{d(1-t)} f(y|0,1,x)^{(1-d)t} f(y|0,0,x)^{(1-d)(1-t)} \\ & \cdot p(1,1,x)^{dt} p(1,0,x)^{d(1-t)} p(0,1,x)^{(1-d)t} p(0,0,x)^{(1-d)(1-t)} \cdot f(x). \end{aligned}$$

Consider the regular sub-model parameterized by $\theta \geq 0$, with the true model indexed by $\theta_0 = 0$,

$$\begin{aligned} f_\theta(y,d,t,x) =& f_\theta(y|1,1,x)^{dt} f_\theta(y|1,0,x)^{d(1-t)} f_\theta(y|0,1,x)^{(1-d)t} f_\theta(y|0,0,x)^{(1-d)(1-t)} \\ & \cdot p_\theta(1,1,x)^{dt} p_\theta(1,0,x)^{d(1-t)} p_\theta(0,1,x)^{(1-d)t} p_\theta(0,0,x)^{(1-d)(1-t)} \\ & \cdot f_\theta(x). \end{aligned}$$

The score function of this sub-model is given by

$$\begin{aligned} s_\theta(y,d,t,x) =& dts_\theta(y|1,1,x) + d(1-t)s_\theta(y|1,0,x) + (1-d)ts_\theta(y|0,1,x) + (1-d)(1-t)s_\theta(y|0,0,x) \\ & + dt\frac{\dot{p}_\theta(1,1,x)}{p_\theta(1,1,x)} + d(1-t)\frac{\dot{p}_\theta(1,0,x)}{p_\theta(1,0,x)} + (1-d)t\frac{\dot{p}_\theta(0,1,x)}{p_\theta(0,1,x)} + (1-d)(1-t)\frac{\dot{p}_\theta(0,0,x)}{p_\theta(0,0,x)} \\ & + t_\theta(x), \end{aligned}$$

where $s_\theta(y|d,t,x) = \partial log f_\theta(y|d,t,x)/\partial\theta$, $\dot{p}_\theta(d,t,x) = \partial p_\theta(d,t,x)/\partial\theta$, and $t_\theta(x) = \partial log f_\theta(x)/\partial\theta$ for $(d,t) \in \mathscr{S}$. For notational simplicity, we suppress subscripts when $\theta = \theta_0$.

Now, the tangent space of this model is characterized by

$$\begin{aligned} \mathscr{T} =& \{dts_{11}(y,x) + d(1-t)s_{10}(y,x) + (1-d)ts_{01}(y,x) + (1-d)(1-t)s_{00}(y,x) \\ & + dtp_{11}(x) + d(1-t)p_{10}(x) + (1-d)tp_{01}(x) + (1-d)(1-t)p_{00}(x) + s(x)\}, \end{aligned}$$

for any functions $\{s_{dt}(\cdot,\cdot), p_{dt}(\cdot)\}_{(d,t)\in\mathscr{S}}$, and $s(\cdot)$ such that, for $(d,t) \in \mathscr{S}$

$$s_{dt}(\cdot,\cdot) \in L_2(\mathscr{Y} \otimes \mathscr{X}), \text{ with } \int s_{dt}(y,x)f(y|d,t,x)dy = 0, \ \forall x \in \mathscr{X}, \tag{3.8.1}$$

$$p_{dt}(\cdot) \in L_2(\mathscr{X}), \text{ with } \sum_{(d,t)\in\mathscr{S}} \int p_{dt}(x)f(x)dx = 0, \tag{3.8.2}$$

and

$$s(\cdot) \in L_2(\mathscr{X}), \text{ with } \int s(x)f(x)dx = 0. \tag{3.8.3}$$

In Step 2, we show that the target parameter associated with the parametric sub-model is *path-wise differentiable*, as defined in Newey (1990).

From Lemma 3.2, we know the ATT can be identified by $\sum_{(d,t)\in\mathscr{S}}(-1)^{d+t}\mathbb{E}[\mathbb{E}[Y|D=d,T=t,X]|D=1,T=1]$ under Assumptions 3.1 and 3.2. For the parameterized sub-model, we define

$$\begin{aligned}\tau(\theta) = & \frac{(\int\int p_\theta(1,1,x)yf_\theta(y|1,1,x)f_\theta(x)dydx - \int\int p_\theta(1,1,x)yf_\theta(y|1,0,x)f_\theta(x)dydx)}{\int p_\theta(1,1,x)f_\theta(x)dx} \\ & - \frac{(\int\int p_\theta(1,1,x)yf_\theta(y|0,1,x)f_\theta(x)dydx - \int\int p_\theta(1,1,x)yf_\theta(y|0,0,x)f_\theta(x)dydx)}{\int p_\theta(1,1,x)f_\theta(x)dx}.\end{aligned} \tag{3.8.4}$$

Note that the derivative of $\tau(\theta)$ with respect to θ, evaluated at $\theta = 0$, is given by

$$\begin{aligned}\left.\frac{d\tau(\theta)}{d\theta}\right|_{\theta=0} = & \sum_{(d,t)\in\mathscr{S}}(-1)^{d+t}\frac{\int\int yp(1,1,x)s(y|d,t,x)f(y|d,t,x)f(x)dydx}{p(1,1)} \\ & + \frac{\int(\tau(x)-\tau)\dot{p}(1,1,x)f(x)dx}{p(1,1)} \\ & + \frac{\int(\tau(x)-\tau)p(1,1,x)t(x)f(x)dx}{p(1,1)}.\end{aligned}$$

For any $w = (y,d,t,x) \in \mathscr{W}$, define

$$\begin{aligned}F_\tau(w) = & \frac{dt(y-m_{1,1}(x))}{p(1,1)} + \frac{p(1,1,x)}{p(1,1)}\left\{-\frac{d(1-t)(y-m_{1,0}(x))}{p(1,0,x)}\right. \\ & \left. - \frac{(1-d)t(y-m_{0,1}(x))}{p(0,1,x)} + \frac{(1-d)(1-t)(y-m_{0,0}(x))}{p(0,0,x)}\right\} \\ & + \frac{dt}{p(1,1)}\sum_{(d,t)\in\mathscr{S}}(-1)^{d+t}\left(m_{d,t}(x) - \int_{\mathscr{X}} m_{d,t}(x)f(x)dx\right).\end{aligned}$$

It can be readily verified that $\left.\frac{d\tau(\theta)}{d\theta}\right|_{\theta=0} = \mathbb{E}[F_\tau(W)s_0(Y,D,T,X)]$, thereby showing $\tau(\theta)$ is path-wise differentiable.

In Step 3, we show that $F_\tau(W)$ is the efficient influence function for τ, which we will accomplish by invoking Theorem 3.1 in Newey (1990). To apply this theorem, we need to verify that $F_\tau(\cdot) \in \mathscr{T}$. By setting

$$\begin{aligned}s_{11}(y,x) &= \frac{y-m_{1,1}(x)}{p(1,1)}, \\ p_{11}(x) &= p(1,1)^{-1}\sum_{(d,t)\in\mathscr{S}}(-1)^{d+t}\left(m_{d,t}(x) - \int_{\mathscr{X}} m_{d,t}(x)f(x)dx\right),\end{aligned}$$

$$s_{dt}(y,x) = (-1)^{d+t}\frac{p(1,1,x)(y-m_{d,t}(x))}{p(d,t,x)p(1,1)},$$

$$p_{dt}(x), s(x) = 0,$$

for $(d,t) \in \mathscr{S}_-$, it is straightforward to show that (3.8.1)–(3.8.3) hold, which leads to the desired result.

Finally, since $p(1,1) = \mathbb{E}\left[I_{d,t}p(1,1,X)p(d,t,X)^{-1}\right]$, for $(d,t) \in \mathscr{S}$, direct manipulation yields that $F_\tau(W) = \eta_{\text{eff}}(W)$. Now, we take the expectation of $\eta^2_{\text{eff}}(W)$ and the semi-parametric efficiency bound follows by standard manipulation. This completes the proof. ■

Proof of Proposition 3.1: The proof follows directly from the LIE as displayed in the main text. ■

Proof of Proposition 3.2:

It follows by Theorem 3.1 that

$$\begin{aligned}\mathbb{E}[\eta_{\text{eff}}(W)^2] &= \frac{1}{\mathbb{E}[DT]^2}\mathbb{E}\left[DT(\tau(Y,X)-\tau)^2 + \sum_{(d,t)\in\mathscr{S}_-}\frac{I_{d,t}p(1,1,X)^2}{p(d,t,X)^2}(Y-m_{d,t}(X))^2\right]\\ &= \frac{1}{\mathbb{E}[DT]^2}\mathbb{E}[DT(\tau(X)-\tau)^2]\\ &\qquad + \mathbb{E}\left[w_{1,1}(D,T)^2(Y-m_{1,1}(X))^2 + \sum_{(d,t)\in\mathscr{S}_-}w_{d,t}(D,T,X,p)^2(Y-m_{d,t}(X))^2\right]\\ &\equiv V_{1,dr} + V_{2,dr},\end{aligned}$$

where the second equality follows from direct manipulations and the fact that

$$\begin{aligned}&\mathbb{E}[DT\cdot(Y-m_{1,1}(X))\cdot(m_{d,t}(X)-\mathbb{E}[m_{d,t}(X)|D=1,T=1])]\\ &\quad = \mathbb{E}[\mathbb{E}[p(1,1,X)\cdot(m_{1,1}(X)-m_{1,1}(X))\cdot(m_{d,t}(X)-\mathbb{E}[m_{d,t}(X)|D=1,T=1])|X]] = 0,\end{aligned}$$

for $(d,t) \in \mathscr{S}$.

Meanwhile, from Part (b) of Proposition 1 in Sant'Anna and Zhao (2020), we have the following decomposition,

$$\mathbb{E}[\eta_{sz}(W)^2] = V_{1,sz} + V_{2,sz},$$

where $V_{1,sz} \equiv \mathbb{E}\left[D(\tau(X)-\tau)^2\right]/p^2$, and

$$V_{2,sz} \equiv \frac{1}{p^2}\mathbb{E}\left[\frac{DT}{\lambda^2}(Y-m_{1,1}(X))^2 + \frac{D(1-T)}{(1-\lambda)^2}(Y-m_{1,0}(X))^2\right.$$

$$+\frac{(1-D)Tp(X)^2}{(1-p(X))^2\lambda^2}(Y-m_{0,1}(X))^2+\frac{(1-D)(1-T)p(X)^2}{(1-p(X))^2(1-\lambda)^2}(Y-m_{0,0}(X))^2\Bigg]. \tag{3.8.5}$$

Under Assumption 3.3, we have that $\mathbb{E}[\mathbb{1}\{T=t\}g(X)]=\mathbb{P}(T=t)\mathbb{E}[g(X)]$, $\mathbb{E}[I_{d,t}Yg(X)]=\mathbb{P}(T=t)\mathbb{E}[\mathbb{1}\{D=d\}Y_t g(X)]$, and $p(d,t,x)=(\mathbb{1}\{t=1\}\lambda+\mathbb{1}\{t=0\}(1-\lambda))p(d,x)$. It then follows that

$$V_{1,dr}=\frac{1}{\lambda p^2}\mathbb{E}[D(\tau(X)-\tau)^2], \tag{3.8.6}$$

and therefore,

$$V_{1,dr}-V_{1,sz}=\frac{1-\lambda}{p^2\lambda}\mathbb{E}[D(\tau(X)-\tau)^2]. \tag{3.8.7}$$

We now focus on $V_{2,dr}$. Observe that

$$\begin{aligned}V_{2,dr}=&\frac{1}{\lambda^2p^2}\Bigg\{\mathbb{E}[DT(Y_1-m_{1,1}(X))^2]+\mathbb{E}\left[\frac{D(1-T)\lambda^2p(X)^2}{(1-\lambda)^2p(X)^2}(Y_0-m_{1,0}(X))^2\right]\\&+\mathbb{E}\left[\frac{(1-D)T\lambda^2p(X)^2}{\lambda^2(1-p(X))^2}(Y_1-m_{0,1}(X))^2\right]+\mathbb{E}\left[\frac{(1-D)(1-T)\lambda^2p(X)^2}{(1-\lambda)^2(1-p(X))^2}(Y_0-m_{0,0}(X))^2\right]\Bigg\}\\=&\frac{1}{p^2}\mathbb{E}\Bigg[\frac{DT}{\lambda^2}(Y-m_{1,1}(X))^2+\frac{D(1-T)}{(1-\lambda)^2}(Y-m_{1,0}(X))^2\\&+\frac{(1-D)Tp(X)^2}{(1-p(X))^2\lambda^2}(Y-m_{0,1}(X))^2+\frac{(1-D)(1-T)p(X)^2}{(1-p(X))^2(1-\lambda)^2}(Y-m_{0,0}(X))^2\Bigg]=V_{2,sz},\end{aligned} \tag{3.8.8}$$

where the first equality follows because $p(d,t,x)=\mathbb{P}(D=d,X=x)\cdot\mathbb{P}(T=t)$ under Assumption 3.3. The desired result then follows from (3.8.7) and (3.8.8). ■

Proof of Lemma 3.1:

Let $\psi_{d,t}(W;w,m)=\mathbb{1}\{dt=1\}w_{1,1}(D,T)Y+\mathbb{1}\{dt\neq1\}\left\{w_{d,t}(D,T,X)(Y-m_{d,t}(X))+w_{1,1}(D,T)m_{d,t}(X)\right\}$, and $\tilde{\tau}_{dr}=\sum_{(d,t)\in\mathscr{S}}(-1)^{d+t}\psi_{d,t}(W;w,m)$. Using $\tilde{\tau}_{dr}$, we decompose $\widehat{\tau}_{dr}$ as

$$\widehat{\tau}_{dr}-\tau=(\widehat{\tau}_{dr}-\tilde{\tau}_{dr})+(\tilde{\tau}_{dr}-\tau). \tag{3.8.9}$$

Note first that the second term, $\tilde{\tau}_{dr}-\tau$, has *i.i.d.* centered summands with bounded variance; thus, it is $O_p(n^{-1/2})$. Now we investigate the behavior of $\widehat{\tau}_{dr}-\tilde{\tau}_{dr}$, for which we make the following decomposition

$$\begin{aligned}\psi_{d,t}(W;\widehat{w},\widehat{m})-\psi_{d,t}(W;w,m)=&(Y-m_{d,t}(X))\left(\widehat{w}_{d,t}-w_{d,t}\right)(W)+m_{d,t}(X)\left(\widehat{w}_{1,1}-w_{1,1}\right)(W)\\&+\left(w_{1,1}-w_{d,t}\right)(W)\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\\&+\left\{\left(\widehat{w}_{1,1}-w_{1,1}\right)(W)-\left(\widehat{w}_{d,t}-w_{d,t}\right)(W)\right\}\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\end{aligned}$$

$$\equiv \Delta_{d,t}^{\psi,1}(W) + \Delta_{d,t}^{\psi,2}(W) + \Delta_{d,t}^{\psi,3}(W),$$

for $(d,t) \in \mathscr{S}$. Here, we use the unifying notation $w_{d,t}(W)$ to denote $w_{d,t}(D,T,X)$ when $(d,t) \in \mathscr{S}_-$ and $w_{1,1}(D,T)$ otherwise. We proceed by establishing convergence rates for each component in the above decomposition.

We first analyze $\Delta_{d,t}^{\psi,1}$. A second-order Taylor expansion of $\psi_{1,1}(W;\widehat{w},\widehat{m})$ around $\mathbb{E}[DT]$ yields that

$$\begin{aligned}
\mathbb{E}_n\left[\Delta_{1,1}^{\psi,1}(W)\right] &= \mathbb{E}_n\left[Y\left(\frac{DT}{\mathbb{E}_n[DT]} - \frac{DT}{\mathbb{E}[DT]}\right)\right] \\
&= -\frac{\mathbb{E}_n[DTY]}{\mathbb{E}[DT]^2} \cdot (\mathbb{E}_n[DT] - \mathbb{E}[DT]) + O_p(|\mathbb{E}_n[DT] - \mathbb{E}[DT]|^2) \\
&= -\frac{\mathbb{E}[DTY]}{\mathbb{E}[DT]^2} \cdot (\mathbb{E}_n[DT] - \mathbb{E}[DT]) + o_p(n^{-1/2}).
\end{aligned} \tag{3.8.10}$$

When $(d,t) \in \mathscr{S}_-$, similar analysis reveals that

$$\begin{aligned}
\mathbb{E}_n\left[\Delta_{d,t}^{\psi,1}(W)\right] =& \mathbb{E}_n\left[(Y - m_{d,t}(X))\left(\widehat{w}_{d,t} - w_{d,t}\right)(W) + m_{d,t}(X)\left(\widehat{w}_{1,1} - w_{1,1}\right)(W)\right] \\
=& \mathbb{E}_n\left[(Y - m_{d,t}(X))\left(\widehat{w}_{d,t} - w_{d,t}\right)(W)\right] \\
& - \frac{\mathbb{E}_n\left[DTm_{d,t}(X)\right]}{\mathbb{E}[DT]^2}(\mathbb{E}_n[DT] - \mathbb{E}[DT]) + O_p(|\mathbb{E}_n[DT] - \mathbb{E}[DT]|^2) \\
=& -\frac{\mathbb{E}\left[DTm_{d,t}(X)\right]}{\mathbb{E}[DT]^2}(\mathbb{E}_n[DT] - \mathbb{E}[DT]) + o_p(n^{-1/2}),
\end{aligned} \tag{3.8.11}$$

where the last equation holds under Assumption 3.4.2(i).

Next, note that $\Delta_{1,1}^{\psi,2}(\cdot) = 0$, and when $(d,t) \in \mathscr{S}_-$, we deduce from Assumption 3.4.2(ii) that

$$\mathbb{E}_n\left[\Delta_{d,t}^{\psi,2}(W)\right] = \mathbb{E}_n\left[(w_{1,1} - w_{d,t})(W)\left(\widehat{m}_{d,t} - m_{d,t}\right)(X)\right] = o_p(n^{-1/2}). \tag{3.8.12}$$

Analogously, $\Delta_{1,1}^{\psi,3}(\cdot)$ is identically zero, and therefore, we only need to focus the other three cases, for which we have

$$\begin{aligned}
&\mathbb{E}_n\left[\Delta_{d,t}^{\psi,3}(W)\right] \\
=& \mathbb{E}_n\left[\left(\left(\widehat{w}_{1,1} - w_{1,1}\right)(W) - \left(\widehat{w}_{d,t} - w_{d,t}\right)(W)\right)\left(\widehat{m}_{d,t} - m_{d,t}\right)(X)\right] \\
=& \mathbb{E}_n\left[\frac{DT}{\mathbb{E}[DT]^2}\left(\widehat{m}_{d,t} - m_{d,t}\right)(X)\right] \cdot (\mathbb{E}_n[DT] - \mathbb{E}[DT]) + O_p(|\mathbb{E}_n[DT] - \mathbb{E}[DT]|^2) && (3.8.13) \\
& - \mathbb{E}_n\left[\left(\widehat{w}_{d,t} - w_{d,t}\right)(W) \cdot \left(\widehat{m}_{d,t} - m_{d,t}\right)(X)\right], && (3.8.14)
\end{aligned}$$

where the second equality follows from a second-order Taylor expansion of $\mathbb{E}_n[DT]$ around $\mathbb{E}[DT]$.

Taking the fact that $\mathbb{E}[DT] > 0$ under Assumption 3.2 (iii) and that $\widehat{m}_{d,t}$ is uniformly convergent to $m_{d,t}$, we obtain

$$\left|\mathbb{E}_n\left[\frac{DT}{\mathbb{E}[DT]^2}\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right]\right| \leq \mathbb{E}_n\left[\left|\frac{DT}{\mathbb{E}[DT]^2}\right| \cdot \left|\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right|\right] \lesssim \left\|\widehat{m}_{d,t}-m_{d,t}\right\|_\infty = o_p(1).$$

Combining this result with $\mathbb{E}_n[DT]-\mathbb{E}[DT] = O_p\left(n^{-1/2}\right)$, we conclude that (3.8.13) is $o_p\left(n^{-1/2}\right)$.

Next, we study $\mathbb{E}_n\left[\left(\widehat{w}_{d,t}-w_{d,t}\right)(W)\cdot\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right]$. Let

$$w^{\dagger}_{d,t}(W) = \frac{I_{d,t}\widehat{p}(1,1,X)}{p(1,1)\widehat{p}(d,t,X)}, \tag{3.8.15}$$

based on which, we have the following decomposition

$$\mathbb{E}_n\left[\left(w^{\dagger}_{d,t}-w_{d,t}\right)(W)\cdot\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right]+\mathbb{E}_n\left[\left(\widehat{w}_{d,t}-w^{\dagger}_{d,t}\right)(W)\cdot\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right] = \Delta^{1,n}_{w,m}+\Delta^{2,n}_{w,m}. \tag{3.8.16}$$

We consider the L_2-norm first. Under Assumption 3.4.2(iii),

$$\Delta^{1,n}_{w,m} = \underbrace{\mathbb{E}\left[\left(w^{\dagger}_{d,t}-w_{d,t}\right)(W)\cdot\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right]}_{\equiv\Delta^{1}_{w,m}}+o_p\left(n^{-1/2}\right).$$

Since $\hat{a}/\hat{b}-a/b = (\hat{a}-a)/b - a(\hat{b}-b)/b^2 - (\hat{a}-a)(\hat{b}-b)/(\hat{b}b)+a(\hat{b}-b)^2/(\hat{b}b^2)$, we have

$$\begin{aligned}
\Delta^{1}_{w,m} =& \mathbb{E}\left[\frac{\delta_{d,t}(W)}{p(d,t,X)}\left(\widehat{p}(1,1,X)-p(1,1,X)\right)\right] \\
&-\mathbb{E}\left[\frac{\delta_{d,t}(W)p(1,1,X)}{p^2(d,t,X)}\left(\widehat{p}(d,t,X)-p(d,t,X)\right)\right] \\
&-\mathbb{E}\left[\frac{\delta_{d,t}(W)}{\widehat{p}(d,t,X)p(d,t,X)}\left(\widehat{p}(1,1,X)-p(1,1,X)\right)\left(\widehat{p}(d,t,X)-p(d,t,X)\right)\right] \\
&+\mathbb{E}\left[\frac{\delta_{d,t}(W)p(1,1,X)}{\widehat{p}(d,t,X)p(d,t,X)^2}\left(\widehat{p}(d,t,X)-p(d,t,X)\right)^2\right] \\
\equiv& \Delta^{1,1}_{w,m}+\Delta^{1,2}_{w,m}+\Delta^{1,3}_{w,m}+\Delta^{1,4}_{w,m},
\end{aligned}$$

where $\delta_{d,t}(W) = p(1,1)^{-1}I_{d,t}\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)$.

For $\Delta_{w,m}^{1,1}$,

$$\begin{aligned}\left|\Delta_{w,m}^{1,1}\right| \leq & p(1,1)^{-1}\left(p_{d,t}^{min}\right)^{-1}\mathbb{E}\left[\left|\left(\widehat{p}(1,1,X)-p(1,1,X)\right)\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right|\right]\\ \leq & O(1)\cdot\left\|\widehat{p}(1,1,\cdot)-p(1,1,\cdot)\right\|_{L_2}\cdot\left\|\widehat{m}_{d,t}-m_{d,t}\right\|_{L_2}\\ = & O_p\left(r_n s_n\right),\end{aligned}$$

where $p_{d,t}^{min}=\inf_{x\in\mathscr{X}}|p(d,t,x)|$. The first inequality holds under Assumption 3.2(iii), and the second one is due to the Cauchy-Schwarz inequality.

Likewise,

$$\begin{aligned}\left|\Delta_{w,m}^{1,2}\right| \leq & p(1,1)^{-1}\sup_{x\in\mathscr{X}}|p(1,1,x)|\left\{\inf_{x\in\mathscr{X}}|p(d,t,x)|\right\}^{-2}\mathbb{E}\left[\left|\left(\widehat{p}(d,t,X)-p(d,t,X)\right)\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right|\right]\\ \leq & O(1)\cdot\left\|\widehat{p}(d,t,\cdot)-p(d,t,\cdot)\right\|_{L_2}\cdot\left\|\widehat{m}_{d,t}-m_{d,t}\right\|_{L_2}\\ = & O_p\left(r_n s_n\right).\end{aligned}$$

To analyze the convergence of the remaining two terms, we can use a similar approach to the one used for the previous two terms. However, to complete the analysis, we need to show that $\widehat{p}(d,t,x)$ is uniformly bounded away from 0 across $\mathscr{X}$, with high probability. Due to the uniform convergence, for any given $\varepsilon\in(0,1/2)$, there is $N_\varepsilon>0$ such that $\sup_{x\in\mathscr{X}}|\widehat{p}(d,t,x)-p(d,t,x)|\leq p_{d,t}^{min}/2$ with probability at least $1-\varepsilon$, whenever $n\geq N_\varepsilon$. Thus, when n is sufficiently large, we have

$$\inf_{x\in\mathscr{X}}|\widehat{p}(d,t,x)|\geq\inf_{x\in\mathscr{X}}|p(d,t,x)|-\sup_{x\in\mathscr{X}}|\widehat{p}(d,t,x)-p(d,t,x)|\geq p_{d,t}^{min}/2>0,$$

with probability $1-\varepsilon$, leading to our desired claim.

The sup-norm case can be handled analogously. Different from the L_2-norm, it is now possible to work directly with the empirical measure, leading to the conclusion that $\Delta_{w,m}^{1,n}=O_p\left(r_n s_n\right)$, without the necessity of imposing Assumption 3.4.2 (iii).

Next,we examine the estimation effect of the normalizing weight as given in $\Delta_{w,m}^{2,n}$. Let $\widehat{p}(1,1)=\mathbb{E}_n\left[I_{d,t}\frac{\widehat{p}(1,1,X)}{\widehat{p}(d,t,X)}\right]$. Again, we focus on L_2-norm first. By definition,

$$\Delta_{w,m}^{2,n}=-\widehat{p}(1,1)^{-1}\cdot\underbrace{\mathbb{E}_n\left[w_{d,t}^{\dagger}(W)\cdot\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)\right]}_{\Delta_{w,m}^{2,1,n}}\cdot\left(\widehat{p}(1,1)-p(1,1)\right).$$

We can further decompose $\Delta_{w,m}^{2,1,n}$ into

$$\begin{aligned}\Delta_{w,m}^{2,1,n} =& \Delta_{w,m}^{1,n} && (3.8.17)\\ &+ (\mathbb{E}_n - \mathbb{E}) \left[w_{d,t}(W) \cdot \left(\widehat{m}_{d,t} - m_{d,t} \right) (X) \right] && (3.8.18)\\ &+ \mathbb{E} \left[w_{d,t}(W) \cdot \left(\widehat{m}_{d,t} - m_{d,t} \right) (X) \right]. && (3.8.19)\\ =& O_p(r_n) + O_p(r_n s_n) + o_p\left(n^{-1/2} \right)\end{aligned}$$

Under Assumptions 3.4.2 (iii, iv), (3.8.17) and (3.8.18) are $O_p(r_n s_n)$ and $o_p\left(n^{-1/2}\right)$, respectively. Since $p_{d,t}(\cdot)$ is uniformly bounded over $\mathscr{X}$, (3.8.19) is $O_p(r_n)$ by the Cauchy-Schwartz inequality.

Analogously, we have

$$\begin{aligned}\widehat{p}(1,1) - p(1,1) =& (\mathbb{E}_n - \mathbb{E}) \left[I_{d,t} \left(\frac{\widehat{p}(1,1,X)}{\widehat{p}(d,t,X)} - \frac{p(1,1,X)}{p(d,t,X)} \right) \right] && (3.8.20)\\ &+ (\mathbb{E}_n - \mathbb{E}) \left[I_{d,t} \frac{p(1,1,X)}{p(d,t,X)} \right] && (3.8.21)\\ &+ \mathbb{E} \left[I_{d,t} \left(\frac{\widehat{p}(1,1,X)}{\widehat{p}(d,t,X)} - \frac{p(1,1,X)}{p(d,t,X)} \right) \right] && (3.8.22)\\ =& O_p(s_n) + O_p\left(n^{-1/2} \right) + o_p\left(n^{-1/2} \right).\end{aligned}$$

Under Assumption 3.4.2 (v), (3.8.20) is $o_p\left(n^{-1/2}\right)$. Since (3.8.21) is a centered *i.i.d.* summand, it is $O_p\left(n^{-1/2}\right)$. Arguing along the same line as for $\Delta_{w,m}^{1}$, we get (3.8.22) is $O_p(s_n)$. Collecting these results, we conclude that both $\Delta_{w,m}^{1,n}$ and $\Delta_{w,m}^{2,n}$ are $O_p(r_n s_n)$.

Once again, analysis under the sup-norm rely directly on empirical measure, thus eliminating the need for conditions on the empirical process. Further details are not provided here for brevity.

To finish the proof of this lemma, we gather the results in (3.8.9), (3.8.10), (3.8.11), (3.8.12), (3.8.14), and (3.8.16), which leads to

$$\begin{aligned}\widehat{\tau}_{dr} - \tau =& \mathbb{E}_n \left[\sum_{(d,t)\in\mathscr{S}} (-1)^{d+t} \psi_{d,t}(W; w, m) - \tau \right] + \tau \left(1 - \frac{\mathbb{E}_n[DT]}{\mathbb{E}[DT]} \right) + O_p(r_n s_n) + o_p\left(n^{-1/2} \right)\\ =& \mathbb{E}_n[\eta_{\text{eff}}(W)] + O_p(r_n s_n) + o_p\left(n^{-1/2} \right).\end{aligned}$$

■

Proof of Theorem 3.2:

We proceed by applying Lemma 3.1. As we are working with the sup-norm, we need to verify the first two conditions in Assumption 3.4.2. Lemmas 3.7 and 3.8 provide the required verification for these conditions. With the bandwidth rate conditions in Assumption 3.5.5 guaranteeing that the leading remainder term is $O_p(r_n s_n) = o_p\left(n^{-1/2}\right)$, we can then derive the asymptotic normality directly from the CLT. ■

Proof of Theorem 3.3:

Proof of Part (a): We have already shown in Theorem 3.2 that $\widehat{\tau}_{dr} - \tau = \mathbb{E}_n[\eta_{\text{eff}}(W)] + o_p\left(n^{-1/2}\right)$. Following a similar line of reasoning, one can easily demonstrate that $\widehat{\tau}_{sz} - \tau = \mathbb{E}_n[\eta_{sz}(W)] + o_p\left(n^{-1/2}\right)$, under Assumptions 3.1, 3.2, 3.5, Condition (i), and the null hypothesis, $\mathbf{H}_0$. Now, by the CLT, we have

$$\sqrt{n}\left(\widehat{\tau}_{dr} - \widehat{\tau}_{sz}\right) \xrightarrow{d} \mathscr{N}\left(0, \mathbb{E}\left[\left(\eta_{\text{eff}}(W) - \eta_{sz}(W)\right)^2\right]\right).$$

It remains to show that

$$\widehat{V}_n \xrightarrow{p} V, \tag{3.8.23}$$

and

$$V = \rho_{sz} > 0. \tag{3.8.24}$$

First, it is implied from the proof of Theorem 3.2 that $\widehat{\eta}_{eff}(w) \xrightarrow{p} \eta_{\text{eff}}(w)$, uniformly in $w \in \mathscr{W}$. In a similar vein, $\widehat{\eta}_{sz}(w) \xrightarrow{p} \eta_{sz}(w)$ uniformly over $\mathscr{W}$, under $\mathbf{H}_0$. Combining these two results, (3.8.23) then follows by the CMT and the weak LLN.

From Proposition 1 in Sant'Anna and Zhao (2020), we know that $\eta_{sz}(\cdot)$ is the efficient influence function for all regular estimators of τ_{sz}, which is equal to τ under $\mathbf{H}_0$. Moreover, since both $\widehat{\tau}_{dr}$ and $\widehat{\tau}_{sz}$ are consistent for τ_{sz} under $\mathbf{H}_0$, it follows from Lemma 2.1 in Hausman (1978) that $\mathbb{E}[\eta_{\text{eff}}(W)\eta_{sz}(W)] = \mathbb{E}[\eta_{sz}(W)^2]$. Hence, $\mathbb{E}\left[\left(\eta_{\text{eff}}(W) - \eta_{sz}(W)\right)^2\right]$ $= \mathbb{E}\left[\eta_{\text{eff}}(W)^2\right] - \mathbb{E}\left[\eta_{sz}(W)^2\right]$. Given this result, (3.8.24) now follows by Proposition 3.2 and the condition that $\mathbb{V}\text{ar}\left[\tau(X)|D=1\right] > 0$.

Proof of Part (b): We proceed by establishing: (i) $\widehat{\tau}_{sz} - \widehat{\tau}_{dr} \xrightarrow{p} \tau_{sz} - \tau_{dr} \neq 0$; (ii) $\widehat{V}_n \xrightarrow{p} V < \infty$, under $\mathbf{H}_1$.

Under Assumption 3.5, and Condition (i) of the theorem, $\widehat{p}(d,t,x) \xrightarrow{p} p(d,t,x)$ and $\widehat{m}_{d,t}(x) \xrightarrow{p} m_{d,t}(x)$, uniformly in x, for $(d,t) \in \mathscr{S}$. Now, applying the LLN, we get $\widehat{\tau}_{dr} \xrightarrow{p} \tau_{dr}$ and $\widehat{\tau}_{sz} \xrightarrow{p} \tau_{sz}$. Result (i) then follows from the CMT. Next, we deduce from the uniform consistency of $\widehat{p}$ and $\widehat{m}$, the CMT, and LLN, that (3.8.23) holds under $\mathbf{H}_1$. Furthermore, Assumptions 3.2(iii) and 3.5.3 ensure that both $\widehat{\eta}_{sz}$ and $\widehat{\eta}_{dr}$ are uniformly bounded, which leads to $V < \infty$. This concludes the proof of part (b). ■

3.8.2 Results on Asymptotic Linear Expansion of Local Polynomial Estimators

In the next subsection, we provide some well-known results about the U-statistics, based on which, we derive uniform stochastic expansions of local polynomial estimators in Section 3.8.2.2.

3.8.2.1 Rates of Convergence: U-Statistics

Let $\{X_i\}_{i=1}^n$ be a random sample from an unknown distribution. Given a real-valued function $h(x_1,...,x_r)$ that possibly depends on the sample size, define

$$U_n = \frac{(n-r)!}{n!} \sum_{s\in S(n,r)} h(X_{s_1},...,X_r),$$

as a r-th order U-statistic with kernel h, where the summation is over $S(n,r)$, the set of permutation $(s_1,...,s_r)$ of size r of the set $\{1,...,n\}$. Since a given function h can always be replaced by a symmetric one, we restrict attention to symmetric kernels in what follows. That is, U_n can be equivalently represented as

$$U_n = \begin{pmatrix} n \\ r \end{pmatrix}^{-1} \sum_{s\in \mathscr{C}(n,r)} h(X_{s_1},...,X_{s_r}),$$

where $\mathscr{C}(n,r)$ is the set of combinations $(s_1,...,s_r)$ of size r of the set $\{1,...,n\}$.

For $1 \leq s \leq r$, define the quantities h_s and σ_s by

$$h_s(x_1,...,x_s) = \mathrm{E}[h(x_1,...,x_s,X_{s+1},...,X_r)] \quad \text{and} \quad \sigma_s = \mathbb{V}\mathrm{ar}\left[h_s(X_1,...,X_s)\right]^{1/2}.$$

We call U_n with kernel h is s^*'th order degenerate if $\sigma_s = 0$ for all $s \leq s^*$.

Lemma 3.3 Let $h : \mathscr{X}^r \to \mathbb{R}$ be a permutation-symmetric, measurable function of r arguments such that $\mathrm{E}[h(X_1,...,X_r)] = 0$, and $\sigma_r < \infty$, then $U_n = O_p\left(\sum_{s=1}^r \frac{\sigma_s}{n^{s/2}}\right)$.

Note that if the U-statistic is s^*-th order degenerate, its convergence rate is $\sum_{s=s^*+1}^r \frac{\sigma_s}{n^{s/2}}$. The lemma follows directly from Markov's inequality, and therefore, we omit the proof.

3.8.2.2 Asymptotic Linear Expansion of Local Polynomial Estimators

In this section, we provide some results on the asymptotic expansion of the local polynomial estimators.

For $(d,t) \in \mathscr{S}_-$, we define the summand of the (local) score function as

$$\widetilde{A}_{d,t}(W,x,\gamma) = \left(I_{d,t} - \frac{\exp(\underline{\mathbf{X}}(x_c)'\gamma_{d,t})}{1+\sum_{(d',t')\in\mathscr{S}_-}\exp(\underline{\mathbf{X}}(x_c)'\gamma_{d',t'})} \right) H(h)\underline{\mathbf{X}}(x_c)\widetilde{K}_{ps}(X;x,h,\lambda),$$

where $H(h)$ is a diagonal matrix with the main diagonal entries being $h^{-|\mathbf{k}|}$, for lexicographic-ordered $\mathbf{k}$, with $0 \leq |\mathbf{k}| \leq p$. Here, we have dropped the subscript of $\underline{\mathbf{X}}$ to ease notational burden. We let $\boldsymbol{\iota}_{-}(\{S_{d,t}\}_{(d,t)\in\mathscr{S}_{-}}) = (S'_{1,0}, S'_{0,1}, S'_{0,0})'$. The local Fisher information matrix evaluated at x can be approximated as

$$\mathscr{I}(x) = diag(\mathbf{p}_{-}(x)) - \mathbf{p}_{-}(x)\mathbf{p}'_{-}(x), \tag{3.8.25}$$

where $\mathbf{p}_{-}(x) = (p(1,0,x), p(0,1,x), p(0,0,x))$. In addition, we define the local hessian as

$$\Sigma^{ps}(x) = \mathbb{E}[\mathscr{I}(X) \otimes H(h)\underline{\mathbf{X}}(x_c)\underline{\mathbf{X}}(x_c)'H(h)\widetilde{K}_{ps}(X; x, h, \lambda)].$$

With these notations in hand, we can introduce several quantities associated with the linear expansion of the PS estimator. For each $(d,t) \in \mathscr{S}_{-}$,

$$\begin{aligned} A_{d,t}(W,x) &= (e_{3,\iota(d,t)} \otimes e_{N_p,1})'\Sigma^{ps}(x)^{-1}\widetilde{\boldsymbol{A}}_{-}(W,x,\gamma^{*}(x)), \\ G^{(ps)}_{d,t}(W,x) &= e'_{3,\iota(d,t)}\mathscr{I}(x)\mathbf{A}_{-}(W,x), \end{aligned}$$

where $\widetilde{\boldsymbol{A}}_{-}(W,x,\gamma) = \boldsymbol{\iota}_{-}(\{\tilde{A}_{d,t}(W,x,\gamma)\}_{(d,t)\in\mathscr{S}_{-}})$, and $\mathbf{A}_{-}(W,x) = \boldsymbol{\iota}_{-}(\{A_{d,t}(W,x)\}_{(d,t)\in\mathscr{S}_{-}})$. For the treated group in $t=1$, let $G^{(ps)}_{1,1}(x) = -\sum_{(d,t)\in\mathscr{S}_{-}} G^{(ps)}_{d,t}(x)$. Additionally, we define, for a given observation X_j

$$\begin{aligned} B^{(ps)}_{n,d,t}(X_j) &= \mathbb{E}[G^{(ps)}_{d,t}(W_i,X_j)|X_j], \\ S^{(ps)}_{n,d,t}(X_j) &= \frac{1}{n-1}\sum_{i\neq j} G^{(ps)}_{d,t}(W_i,X_j) - \mathbb{E}[G^{(ps)}_{d,t}(W_i,X_j)|X_j], \\ R^{(ps)}_{n,d,t}(X_j) &= \hat{p}(d,t,X_j) - p(d,t,X_j) - B^{(ps)}_{n,d,t}(X_j) - S^{(ps)}_{n,d,t}(X_j). \end{aligned} \tag{3.8.26}$$

The three quantities represent the bias, the first-order stochastic part, and the remaining terms derived from the decomposition of the PS estimator, respectively.

Focusing on the OR model, for $(d,t) \in \mathscr{S}$, the leave-one-out local polynomial estimator has a closed-form solution given by

$$\widehat{m}_{d,t}(X_j) = \frac{1}{n-1}\sum_{i\neq j} e'_{N_q,1}\hat{\Sigma}^{or\;-1}_{d,t}(X_j)\underline{\mathbf{X}}_i(X_j)H(b_{d,t})I_{d,t,i}Y_i\widetilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}),$$

where $\widehat{\Sigma}^{or}_{d,t}(X_j) = \frac{1}{n-1}\sum_{i\neq j} I_{d,t,i}H(b_{d,t})\underline{\mathbf{X}}_i(x_c)\underline{\mathbf{X}}_i(x_c)'H(b_{d,t})\widetilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t})$.

Analogous to the PS case, we use $B^{(or)}_{n,d,t}$, $S^{(or)}_{n,d,t}$, and $R^{(or)}_{n,d,t}$ to represent the bias, the first-order stochastic and the remainder terms, respectively. For a given observation X_j, these quantities are specified as

$$\begin{aligned}
B^{(or)}_{n,d,t}(X_j) &= \mathbb{E}[G^{(or)}_{d,t}(W_i,X_j)|X_j], \\
S^{(or)}_{n,d,t}(X_j) &= \frac{1}{n-1}\sum_{i\neq j} G^{(or)}_{n,d,t}(W_i,X_j) - \mathbb{E}[G^{(or)}_{d,t}(W_i,X_j)|X_j], \\
R^{(or)}_{n,d,t}(X_j) &= \hat{m}_{d,t}(X_j) - m_{d,t}(X_j) - B^{(or)}_{n,d,t}(X_j) - S^{(or)}_{n,d,t}(X_j),
\end{aligned}$$

where

$$\begin{aligned}
&G^{(or)}_{d,t}(W_i,X_j) = e'_{N_q,1}\Sigma^{or}_{d,t}(X_j)^{-1}H(b_{d,t})\underline{\mathbf{X}}_i(X_j)I_{d,t,i}\xi^{or}_{d,t,i}(X_j)\widetilde{K}_{or}(X_i;X_j,b_{d,t},\vartheta_{d,t}), \\
&\Sigma^{or}_{d,t}(x) = \mathbb{E}[I_{d,t,i}H(b_{d,t})\underline{\mathbf{X}}_i(x_c)\underline{\mathbf{X}}_i(x_c)'H(b_{d,t})\widetilde{K}_{or}(X;X_j,b_{d,t},\vartheta_{d,t})], \\
&\xi^{or}_{d,t}(x) = I_{d,t}(Y - \underline{\mathbf{X}}(x)'\beta^*_{d,t}).
\end{aligned}$$

Lemma 3.4 Suppose Assumptions 3.1, 3.2, and 3.5 are satisfied. In addition, Assumptions 3.5.2 (ii) and 3.5.5 (iv)–(vii) hold for $(d,t) = (1,1)$. Then, for $(d,t) \in \mathscr{S}$,

$$\sup_{j\in\mathbb{N}_n}\left|B^{(ps)}_{n,d,t}(X_j)\right| = O_p(h^{p+1}+\lambda_o+\lambda_u), \tag{3.8.27}$$

$$\sup_{j\in\mathbb{N}_n}\left|S^{(ps)}_{n,d,t}(X_j)\right| = O_p\left(\sqrt{\log n/\left(nh^{\upsilon_c}\right)}\right), \tag{3.8.28}$$

$$\sup_{j\in\mathbb{N}_n}\left|R^{(ps)}_{n,d,t}(X_j)\right| = O_p\left(\left(h^{p+1}+\lambda_o+\lambda_u+\sqrt{\log n/\left(nh^{\upsilon_c}\right)}\right)^2\right), \tag{3.8.29}$$

$$\begin{aligned}
\sup_{j\in\mathbb{N}_n}\left|B^{(or)}_{n,d,t}(X_j)\right| &= O_p(b^{q+1}_{d,t}+\vartheta_{d,t,o}+\vartheta_{d,t,u}), \\
\sup_{j\in\mathbb{N}_n}\left|S^{(or)}_{n,d,t}(X_j)\right| &= O_p\left(\sqrt{\log n/\left(n{b_{d,t}}^{\upsilon_c}\right)}\right), \\
\sup_{j\in\mathbb{N}_n}\left|R^{(or)}_{n,d,t}(X_j)\right| &= O_p\left(\left(b^{p+1}_{d,t}+\vartheta_{d,t,o}+\vartheta_{d,t,u}+\sqrt{\log n/\left(n{b_{d,t}}^{\upsilon_c}\right)}\right)^2\right).
\end{aligned}$$

Before stating the proof, we need to introduce some additional notations. Since kernel functions K and L are supported on $[-1,1]^{\upsilon_c}$, the effective support of $K((\cdot - x_c)/h)$ is $\mathscr{S}_{x_c,h} = \{z : x_c + hz \in \mathscr{X}\} \cap [-1,1]^{\upsilon_c}$. When $\mathscr{S}_{x_c,h} = [-1,1]^{\upsilon_c}$, x is an interior point, otherwise x lies close to the boundary. For any measurable set $\mathscr{S} \subset [-1,1]^{\upsilon_c}$, let $\nu_{\mathbf{k}}(\mathscr{S}) = \int_{\mathscr{S}} \mathbf{u}^{\mathbf{k}}K(\mathbf{u})d\mathbf{u}$ and $\varkappa_{\mathbf{k}}(\mathscr{S}) = \int_{\mathscr{S}} \mathbf{u}^{\mathbf{k}}K^2(\mathbf{u})d\mathbf{u}$. Now we let the $N_\ell \times N_\ell$ matrices $\mathbf{Q}_\ell(x_c)$ and $\mathbf{T}_\ell(x_c)$, and the

$N_\ell \times n_k$ matrix $\mathbf{M}_{\ell,k}(x_c)$ be defined as

$$\mathbf{Q}_\ell(x_c) = \begin{pmatrix} \mathbf{Q}^{(0,0)}(\mathscr{S}_{x_c,h}) & \cdots & \mathbf{Q}^{(0,\ell)}(\mathscr{S}_{x_c,h}) \\ \vdots & \ddots & \vdots \\ \mathbf{Q}^{(\ell,0)}(\mathscr{S}_{x_c,h}) & \ldots & \mathbf{Q}^{(\ell,\ell)}(\mathscr{S}_{x_c,h}) \end{pmatrix}, \tag{3.8.30}$$

$$\mathbf{T}_\ell(x_c) = \begin{pmatrix} \mathbf{T}^{(0,0)}(\mathscr{S}_{x_c,h}) & \cdots & \mathbf{T}^{(0,\ell)}(\mathscr{S}_{x_c,h}) \\ \vdots & \ddots & \vdots \\ \mathbf{T}^{(\ell,0)}(\mathscr{S}_{x_c,h}) & \ldots & \mathbf{T}^{(\ell,\ell)}(\mathscr{S}_{x_c,h}) \end{pmatrix},$$

$$\mathbf{M}_{\ell,k}(x_c) = \begin{pmatrix} \mathbf{Q}^{(0,k)}(\mathscr{S}_{x_c,h}) \\ \cdots \\ \mathbf{Q}^{(\ell,k)}(\mathscr{S}_{x_c,h}) \end{pmatrix},$$

where $\mathbf{Q}_\ell^{(i,j)}(\mathscr{S})$ and $\mathbf{T}_\ell^{(i,j)}(\mathscr{S})$ are $n_i \times n_j$ matrices with their respective (l,m)-th element given by $v_{\pi_i(l)+\pi_j(m)}(\mathscr{S})$ and $\varkappa_{\pi_i(l)+\pi_j(m)}(\mathscr{S})$. When x is a boundary point, these quantities are not invariant to x, and thus, capture the boundary effects.

Proof of Lemma 3.4:

Given that our data is a random sample, it is straightforward to show the “leave-one-out” estimators considered in the lemma is asymptotically equivalent to the usual “leave-in” estimators. See Rothe and Firpo (2019) for a detailed exposition. We therefore focus on the “leave-in” versions in what follows.

We prove the results for PS only. The case for OR follows by generalizing Proposition 7 of Fan and Guerre (2016) to the case where discrete covariates are accommodated. This generalization can be achieved by employing the techniques similar to those presented here.

For (3.8.27), we have

$$\begin{aligned} \sup_{x\in\mathscr{X}} \left\| B_{n,d,t}^{(ps)}(x) \right\| &= \sup_{x\in\mathscr{X}} \left\| e'_{3,\iota(d,t)} \mathscr{I}(x)(I_3 \otimes e'_{N_p,1}) \Sigma^{ps}(x)^{-1}\, \mathbb{E}[\widetilde{\boldsymbol{A}}_-(W,x,\gamma^*(x))] \right\| \\ &\leq \sup_{x\in\mathscr{X}} \left\| e'_{3,\iota(d,t)} \mathscr{I}(x) \right\| \cdot \sup_{x\in\mathscr{X}} \left\| (I_3 \otimes e'_{N_p,1}) \Sigma^{ps}(x)^{-1} \right\| \cdot \sup_{x\in\mathscr{X}} \left\| \mathbb{E}[\widetilde{\boldsymbol{A}}_-(W,x,\gamma^*(x))] \right\|. \end{aligned}$$

By definition, $\sup_{x\in\mathscr{X}} \|\mathscr{I}(x)\| = O(1)$. Standard change of variable gives

$$\Sigma^{ps}(x) = \mathscr{I}(x) \otimes \mathbf{Q}_p(x_c) f_X(x) + O\left(h + \lambda_o + \lambda_u\right). \tag{3.8.31}$$

Since $\inf_{x\in\mathscr{X}}\lambda_{min}(\mathscr{I}(x)\otimes\mathbf{Q}_p(x_c)) = \inf_{x\in\mathscr{X}}\lambda_{min}(\mathscr{I}(x))\cdot\inf_{x_c\in\mathscr{X}_c}\lambda_{min}(\mathbf{Q}_p(x_c)) > 0$ and $\inf_{x\in\mathscr{X}} f_X(x) > 0$ under Assumptions 3.2(iii), 3.5.6, and 3.5.1, we get

$$\sup_{x\in\mathscr{X}}\left\|\mathscr{I}(x)^{-1}\otimes\mathbf{Q}_p(x_c)^{-1}\cdot f_X(x)^{-1}\right\| = O(1), \tag{3.8.32}$$

and thus, $\sup_{x\in\mathscr{X}}\left\|\Sigma^{ps}(x)^{-1}\right\| = O(1)$. Now, from Lemma 3.5, we conclude that $\sup_{x\in\mathscr{X}}\left\|B^{(ps)}_{n,d,t}(x)\right\| = O(h^{p+1}+\lambda_o+\lambda_u)$.

Having just demonstrated that $\Sigma^{ps}(x)^{-1}$ is uniformly bounded over $\mathscr{X}$, we can now apply Lemma 3.5 and the CMT to deduce (3.8.28).

To establish (3.8.29), the proof proceed through three steps. First, we demonstrate the existence of a global maximizer for the local log-likelihood function defined in (3.3.6). Subsequently, we obtain the uniform asymptotic linear expansion for the local maximum likelihood estimator. Finally, we apply the delta method to verify that the remainder term exhibits the required rate.

Step 1: Define $\bar{\gamma} = (I_3\otimes H(h)^{-1})\gamma$ and $\bar{\gamma}^*(\cdot) = (I_3\otimes H(h)^{-1})\gamma^*(\cdot)$. Using the scaled parameters, we rewrite the likelihood as

$$\begin{aligned}\mathscr{L}^{ps}_n(\bar{\gamma};x) =& \frac{1}{n}\sum_{i=1}^{n}\sum_{(d',t')\in\mathscr{S}_-} I_{d,t}H(h)\underline{\mathbf{X}}(x_c)'\bar{\gamma}_{d,t}\\ &-\log\left(1+\sum_{(d',t')\in\mathscr{S}_-}\exp\left(H(h)\underline{\mathbf{X}}(x_c)'\bar{\gamma}_{d',t'}\right)\right)\widetilde{K}_{ps}(X_i;x,h,\lambda).\end{aligned} \tag{3.8.33}$$

The gradient and hessian of $\mathscr{L}^{ps}_n(\bar{\gamma};x)$ with respect to $\bar{\gamma}$ are given by

$$\nabla_{\bar{\gamma}}\mathscr{L}^{ps}_n(\bar{\gamma};x) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}_-(W_i,x,\gamma), \qquad \nabla^2_{\bar{\gamma}\bar{\gamma}'}\mathscr{L}^{ps}_n(\bar{\gamma};x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(W_i,x,\gamma),$$

where

$$\begin{aligned}&\mathbf{H}(X,x,\gamma) = \mathscr{I}(X_c,x_c,\gamma)\otimes\widetilde{H}(X,x,h,\lambda),\\ &\widetilde{H}(X,x,h,\lambda) = H(h)\underline{\mathbf{X}}(x_c)\underline{\mathbf{X}}(x_c)'H(h)\widetilde{K}_{ps}(X;x,h,\lambda),\\ &\mathscr{I}(X_c,x_c,\gamma) = diag(\boldsymbol{\Psi}_-(X_c,x_c,\gamma)) - \boldsymbol{\Psi}_-(X_c,x_c,\gamma)\boldsymbol{\Psi}_-(X_c,x_c,\gamma)',\\ &\boldsymbol{\Psi}_-(X,x,\gamma) = \boldsymbol{\iota}_-(\{\Psi_{d,t}(\underline{\mathbf{X}}(x),\gamma)\}_{(d,t)\in\mathscr{S}_-}),\\ &\Psi_{d,t}(x,\gamma) = \frac{\exp\left(x'\gamma_{d,t}\right)}{1+\sum_{(d',t')\in\mathscr{S}_-}\exp\left(x'\gamma_{d',t'}\right)}.\end{aligned}$$

Next, we define the following two events

$$E_{1n}(c) = \left\{ \sup_{x\in\mathscr{X}} \left\| \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{A}}_{-}(W_i, x, \gamma^*(x)) \right\| < c\kappa_n \right\},$$

$$E_{2n}(c) = \left\{ \inf_{x\in\mathscr{X}} \lambda_{min}\left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{H}(X;x,h,\lambda) \right) > c \right\},$$

for $c > 0$ and $\kappa_n = \sqrt{\log n/(nh^{\upsilon_c})} + h^{p+1} + \lambda_u + \lambda_o$.

By Lemma 3.5, we deduce that $\mathbb{P}(E_{1n}(c_1)) \to 1$, for any fixed $c_1 > 0$.

Now, standard change-of-variable analysis gives

$$\mathbb{E}[\widetilde{H}(X;x,h,\lambda)] = \mathbf{Q}_p(x_c) f_X(x) + O\left(h + \lambda_o + \lambda_u\right).$$

Under Assumptions 3.5.1 and 3.5.6, $\inf_{x\in\mathscr{X}} f_X(x) > 0$ and $\inf_{x_c\in\mathscr{X}_c} \lambda_{min}\left(\mathbf{Q}_p(x_c)\right) > 0$. As a result, there exists $c_2 > 0$ such that $\inf_{x\in\mathscr{X}} \lambda_{min}(\mathbb{E}[\widetilde{H}(X;x,h,\lambda)]) \geq c_2$, when n is sufficiently large. Coupled with the fact that

$$\sup_{x\in\mathscr{X}} \left\| \frac{1}{n}\sum_{i=1}^{n} \widetilde{H}(X_i;x,h,\lambda) - \mathbb{E}[\widetilde{H}(X;x,h,\lambda)] \right\| = O_p\left(\sqrt{\log n/(nh^{\upsilon_c})} \right).$$

which is a consequence of Lemma 5 from Fan and Guerre (2016), we deduce that $\mathbb{P}(E_{2n}(c)) \to 1$, for $c \leq c_2$.

Next, we define a neighborhood of $\bar{\gamma}^*(\cdot)$,

$$\Gamma(\delta) = \{\gamma(\cdot) : \|\bar{\gamma}(\cdot) - \bar{\gamma}^*(\cdot)\|_\infty \leq \delta\kappa_n\}.$$

Theorem 1 in Tanabe and Sagae (1992) implies that

$$\inf_{x,y\in\mathscr{X}} \mathscr{I}(x,y,\gamma(y)) > \inf_{x,y\in\mathscr{X}} \left\{ \prod_{(d,t)\in\mathscr{S}_-} \Psi_{d,t}(\{\underline{\mathbf{x}}(y)'\gamma_{d,t}(y)\}_{(d,t)\in\mathscr{S}_-}) \right\} \cdot I_3, \tag{3.8.34}$$

in the sense that their difference is positive definite. For any $\delta > 0$, if $\gamma \in \Gamma(\delta)$, Assumption 3.5.5(ii) implies that $\|\gamma(\cdot) - \gamma^*(\cdot)\|_\infty = o(1)$. This further suggests that, when n is sufficiently large, the right-hand side of (3.8.34) is bounded from below by $c_3 I_3$, for some positive constant c_3.

The analysis leading up to this point demonstrates that for for a given $c_1 > 0$, it is possible to select n large enough such that $\mathbb{P}(E_{1n}(c_1)) > 1 - \varepsilon/2$, $\mathbb{P}(E_{2n}(c_2)) > 1 - \varepsilon/2$, and (3.8.34) is satisfied. Now, set $\delta_0 > 2c_1 c_2^{-1} c_3^{-1}$. Then, for any $\gamma(\cdot) \in \partial\Gamma(\delta_0)$, i.e., $\|\bar{\gamma}(x) - \bar{\gamma}^*(x)\| = \delta_0\kappa_n$, for all $x \in \mathscr{X}$, we have $\sup_{x\in\mathscr{X}}\left\{\mathscr{L}_n^{ps}(\bar{\gamma}(x);x) - \mathscr{L}_n^{ps}(\bar{\gamma}^*(x);x)\right\} < 0$, with a probability of at least $1-\varepsilon$. This is because

$$\begin{aligned}
&\sup_{x\in\mathscr{X}}\{\mathscr{L}_n^{ps}(\bar{\gamma}(x);x)-\mathscr{L}_n^{ps}(\bar{\gamma}^*(x);x)\}\\
=&\sup_{x\in\mathscr{X}}\left\{\nabla_{\bar{\gamma}}\mathscr{L}_n^{ps}(\bar{\gamma}^*(x);x)(\bar{\gamma}-\bar{\gamma}^*(x))-(\bar{\gamma}(x)-\bar{\gamma}^*(x))'\left(-\nabla^2_{\bar{\gamma}\bar{\gamma}'}\mathscr{L}_n^{ps}(\bar{\gamma}^\dagger;x)\right)(\bar{\gamma}(x)-\bar{\gamma}^*(x))/2\right\}\\
\leq&\left(\sup_{x\in\mathscr{X}}\left\|\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{A}}_-(W_i,x,\gamma^*(x))\right\|-c_1\kappa_n\right)\cdot\delta_0\kappa_n\\
<&0,
\end{aligned}$$

where $\bar{\gamma}^\dagger$, dependent on x, lies between $\bar{\gamma}(x)$ and $\bar{\gamma}^*(x)$. Since $\mathscr{L}_n^{ps}(\bar{\gamma};x)$ is continuous, a local maximum, denoted by $\hat{\bar{\gamma}}(x)$, exists within the compact set $\{\bar{\gamma}:\|\bar{\gamma}-\bar{\gamma}^*(x)\|\leq\delta_0\kappa_n\}$, for any $x\in\mathscr{X}$. Furthermore, due to the concavity of $\mathscr{L}_n^{ps}(\cdot;x)$, $\hat{\bar{\gamma}}(x)$ maximizes $\mathscr{L}_n^{ps}(\cdot;x)$ over $\mathbb{R}^{3N_p}$ for any $x\in\mathscr{X}$. Hence, $\hat{\bar{\gamma}}(\cdot)$ is the global maximizer of $\mathscr{L}_n^{ps}(\bar{\gamma}(\cdot);\cdot)$ with a probability exceeding $1-\varepsilon$. As ε is arbitrary and δ_0 is independent of x, it can be inferred that $\left\|\hat{\bar{\gamma}}(\cdot)-\bar{\gamma}^*(\cdot)\right\|_\infty=O_p(\kappa_n)$.

Step 2: We proceed to derive the uniform asymptotic linear expansion of $\hat{\bar{\gamma}}(\cdot)-\bar{\gamma}^*(\cdot)$. Expanding $\mathscr{L}_n^{ps}(\bar{\gamma};x)$ using a third-order Taylor series and rearranging the terms lead to

$$\hat{\bar{\gamma}}(x)-\bar{\gamma}^*(x)=\frac{1}{n}\sum_{i=1}^n\Sigma^{ps}(x)^{-1}\widetilde{\boldsymbol{A}}_-(W_i,x,\gamma^*(x))+R^\gamma(X_j),$$

where

$$\begin{aligned}
R^\gamma(x)=&-\left(\Sigma_n^{ps}(x)^{-1}-\Sigma^{ps}(x)^{-1}\right)\cdot\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{A}}_-(W_i,x,\gamma^*(x))-\Sigma_n^{ps}(x)^{-1}\mathbf{C}_n(x),\\
\mathbf{C}_n(x)=&\frac{1}{2n}\sum_{i=1}^n\sum_{(d,t)\in\mathscr{S}_-}\sum_{(d',t')\in\mathscr{S}_-}(\hat{\bar{\gamma}}_{d,t}(x)-\bar{\gamma}^*_{d,t}(x))'H(h)\underline{\mathbf{X}}_i(x_c)\underline{\mathbf{X}}_i(x_c)'H(h)(\hat{\bar{\gamma}}_{d',t'}(x)-\bar{\gamma}^*_{d',t'}(x))\\
&\cdot\dot{\mathscr{I}}_{\iota(d,t),\iota(d',t')}(X_{c,i},x_c,\tilde{\gamma})\otimes\underline{\mathbf{X}}_i(x_c)H(h)\widetilde{K}_{ps}(X_i;x,h,\lambda),
\end{aligned}$$

for an intermediate point $\tilde{\gamma}$ lying between $\widehat{\gamma}(x)$ and $\gamma^*(x)$, $\Sigma_n^{ps}(\cdot)=\frac{1}{n}\sum_{i=1}^n\mathbf{H}(W_i,\cdot,\gamma^*(\cdot))$, and

$$\begin{aligned}
&\dot{\mathscr{I}}_{\iota(d_1,t_1),\iota(d_2,t_2)}=\boldsymbol{\iota}_-\left(\left\{\dot{\mathscr{I}}^{(d_3,t_3)}_{\iota(d_1,t_1),\iota(d_2,t_2)}\right\}_{(d_3,t_3)\in\mathscr{S}_-}\right),\\
&\dot{\mathscr{I}}^{(d_3,t_3)}_{\iota(d_1,t_1),\iota(d_2,t_2)}(X_c,x_c,\gamma)=\mathbb{1}\{(d_1,t_1)=(d_2,t_2)\}\Psi_{d_1,t_1}(\underline{\mathbf{X}}(x_c),\gamma)(\mathbb{1}\{(d_1,t_1)=(d_3,t_3)\}-\Psi_{d_3,t_3}(\underline{\mathbf{X}}(x_c),\gamma))\\
&+\sum_{\ell_1,\ell_2\in\{1,2\},\ell_1\neq\ell_2}\Psi_{d_{\ell_1},t_{\ell_1}}(\underline{\mathbf{X}}(x_c),\gamma)\Psi_{d_{\ell_2},t_{\ell_2}}(\underline{\mathbf{X}}(x_c),\gamma)(\mathbb{1}\{(d_{\ell_2},t_{\ell_2})=(d_3,t_3)\}-\Psi_{d_3,t_3}(\underline{\mathbf{X}}(x_c),\gamma)).
\end{aligned}$$

In view of (3.8.31) and (3.8.32), $\left\|\Sigma^{ps}(\cdot)^{-1}\right\| = O(1)$. Taking this into account, along with Lemma 3.5, we obtain

$$\begin{aligned}
&\sup_{x\in\mathscr{X}}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_{-}(W_i,x,\gamma^*)-\mathbb{E}[\mathbf{A}_{-}(W,x,\gamma^*)]\right\| = O_p\left(\sqrt{\log n/(nh^{\upsilon_c})}\right),\\
&\sup_{x\in\mathscr{X}}\left\|\mathbb{E}[\mathbf{A}_{-}(W,x,\gamma^*)]\right\| = O_p\left(h^{p+1}+\lambda_o+\lambda_u\right).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sup_{x\in\mathscr{X}}\left\|\Sigma_n^{ps}(x)^{-1}-\Sigma^{ps}(x)^{-1}\right\|\\
&\quad\leq \sup_{x\in\mathscr{X}}\|\Sigma_n^{ps}(x)\|^{-1}\cdot\sup_{x\in\mathscr{X}}\|\Sigma_n^{ps}(x)-\Sigma^{ps}(x)\|\cdot\sup_{x\in\mathscr{X}}\|\Sigma^{ps}(x)\|^{-1}\\
&\quad= O_p(1)\cdot O_p\left(\sqrt{\log n/(nh^{\upsilon_c})}\right)\cdot O(1)\\
&\quad= O_p\left(\sqrt{\log n/(nh^{\upsilon_c})}\right).
\end{aligned}$$

where the first inequality is a result of the relationship $A^{-1}-B^{-1}=-A^{-1}(A-B)B^{-1}$ and the Cauchy-Schwarz inequality. The next line is derived from (3.8.31) and (3.8.32), and arguments similar to those employed in the proof of Lemma 5 in Fan and Guerre (2016).

By the triangular inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sup_{x\in\mathscr{X}}\|\mathbf{C}_n(x)\| \leq& \frac{1}{2n}\sum_{i=1}^{n}\sum_{(d,t)\in\mathscr{S}_{-}}\sum_{(d',t')\in\mathscr{S}_{-}}\left\|\dot{\mathscr{I}}_{\iota(d,t),\iota(d',t')}(X_{c,i},x_c,\tilde{\gamma}(x))\right\|\\
&\cdot\left\|\hat{\bar{\gamma}}_{d,t}(x)-\bar{\gamma}^*_{d,t}(x)\right\|\cdot\left\|\hat{\bar{\gamma}}_{d',t'}(x)-\bar{\gamma}^*_{d',t'}(x)\right\|\cdot\|H(h)\underline{\mathbf{X}}_i(x_c)\|^3\cdot\left|\widetilde{K}_{ps}(X_i;x,h,\lambda)\right|\\
\lesssim& \max_{(d,t),(d',t')\in\{0,1\}}\sup_{x,z\in\mathscr{X}}\left\{\left\|\dot{\mathscr{I}}_{\iota(d,t),\iota(d',t')}(z_c,x_c,\tilde{\gamma}(x))\right\|\cdot\left\|\hat{\bar{\gamma}}_{d,t}(x)-\bar{\gamma}^*_{d,t}(x)\right\|\cdot\left\|\hat{\bar{\gamma}}_{d',t'}(x)-\bar{\gamma}^*_{d',t'}(x)\right\|\right\}
\end{aligned}\tag{3.8.35}$$

$$\cdot\frac{1}{n}\sum_{i=1}^{n}\sup_{x\in\mathscr{X}}\left\{\left|K_h^{ps}(\underline{\mathbf{X}}_i^{(1)}(x_c))\right|\cdot\|H(h)\underline{\mathbf{X}}_i(x_c)\|^3\right\}.\tag{3.8.36}$$

When $\tilde{\gamma}$ converges uniformly to γ^*, as established in the first step, $\left\|\dot{\mathscr{I}}_{\iota(d,t),\iota(d',t')}(z_c,x_c,\tilde{\gamma}(x))\right\|$ in (3.8.35) is asymptotically bounded, uniformly in $x,z\in\mathscr{X}$, and for each $(d,t),(d',t')\in\mathscr{S}_{-}$. In addition, we can deduce from a standard change of variable argument that (3.8.36) is $O_p(1)$. Hence, it can be concluded that $\sup_{x\in\mathscr{X}}\|\mathbf{C}_n(x)\| = O_p\left(\kappa_n^2\right)$. As a result, we obtain $\sup_{x\in\mathscr{X}}\|R^{\gamma}(x)\| = O_p\left(\kappa_n^2\right)$.

Step 3: We note that $\widehat{p}(d,t,x)-p(d,t,x) = \Psi_{d,t}(e_{N_p,1},\widehat{\gamma}(x))-\Psi_{d,t}(e_{N_p,1},\gamma^*(x))$ and $\nabla_{\gamma_{d,t}}\Psi_{d,t}(e_{N_p,1},\gamma^*(x)) = e'_{3,\iota(d,t)}\mathscr{I}(x)$. Utilizing the delta method in conjunction with the uniform expansion obtained in Step 2 then establishes (3.8.29). This completes the proof of the lemma. ■

Lemma 3.5 Suppose that the conditions of Lemma 3.4 hold. Then

$$\sup_{x\in\mathscr{X}}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}_{-}(W_i,x,\gamma^*(x))-\mathbb{E}[\widetilde{\boldsymbol{A}}_{-}(W,x,\gamma^*(x))]\right\| = O_p\left(\left(\log n/\left(nh^{\upsilon_c}\right)\right)^{1/2}\right), \tag{3.8.37}$$

$$\sup_{x\in\mathscr{X}}\left\|\mathbb{E}[\widetilde{\boldsymbol{A}}_{-}(W,x,\gamma^*(x))]\right\| = O\left(h^{p+1}+\lambda_o+\lambda_u\right). \tag{3.8.38}$$

Proof of Lemma 3.5:

The proof of (3.8.37) proceeds along similar lines as in Lemma 5 of Fan and Guerre (2016). For any given vector $\mathbf{k}$ with $0\le|\mathbf{k}|\le p$, define

$$\widetilde{A}_{d,t}^{(\mathbf{k})}(W,x,\gamma) = \left(I_{d,t}-\Psi_{d,t}(\underline{\mathbf{X}}(x_c),\gamma)\right)h^{-|\mathbf{k}|}(X_c-x_c)^{\mathbf{k}}\widetilde{K}(X;x,h,\lambda),$$
$$\widetilde{A}_{d,t}^{\dagger,(\mathbf{k})}(W,x_c,\gamma) = \left(I_{d,t}-\Psi_{d,t}(\underline{\mathbf{X}}(x_c),\gamma)\right)h^{-|\mathbf{k}|}(X_c-x_c)^{\mathbf{k}}K\left(\frac{X_c-x_c}{h}\right),$$

for $(d,t)\in\mathscr{S}_{-}$, and let $\kappa_n = \left(\log n/\left(nh^{\upsilon_c}\right)\right)^{1/2}$. Assumption 3.5.5 implies that $\kappa_n\to 0$. Moreover, under Assumptions 3.5.1, 3.5.2, and 3.5.4, we have that, for any $\varepsilon>0$, there exists $\delta_n = n^{-\kappa_a}$ such that (i)

$$\max_{i\in\mathbb{N}_n}\left|\widetilde{A}_{d,t}^{\dagger,(\mathbf{k})}(W_i,x_c,\gamma^*(x))-\widetilde{A}_{d,t}^{\dagger,(\mathbf{k})}(W_i,x_c',\gamma^*(x'))\right|\le h^{\upsilon_c}\kappa_n\varepsilon/3, \tag{3.8.39}$$

$$\left|\mathbb{E}\left[\widetilde{A}_{d,t}^{\dagger,(\mathbf{k})}(W,x_c,\gamma^*(x))\right]-\mathbb{E}\left[\widetilde{A}_{d,t}^{\dagger,(\mathbf{k})}(W,x_c',\gamma^*(x'))\right]\right|\le h^{\upsilon_c}\kappa_n\varepsilon/3, \tag{3.8.40}$$

for $(d,t)\in\mathscr{S}_{-}$ and for all $x,x'\in\mathscr{X}$ such that $x_d = x_d'$ and $\|x_c-x_c'\|\le\delta_n$; (ii) there is a positive integer $J_n = O(n^{\kappa_b})$, $\kappa_b>0$, and a set $\{x_j\}_{j=1}^{J_n}\subset\mathscr{X}$, such that for all $x\in\mathscr{X}$, there exists a j satisfying $x\in\mathscr{B}(x_j,\delta_n)\cap\mathscr{X}$, and for all $x'\in\mathscr{B}(x_j,\delta_n)$, $x_d' = x_{d,j}$. As a result, $\mathscr{X} = \bigcup_{j=1}^{J_n}\left(\mathscr{B}(x_j,\delta_n)\cap\mathscr{X}\right)$.

Now, observe that, for $(d,t)\in\mathscr{S}_{-}$

$$\begin{aligned}
&\sup_{x\in\mathscr{X}}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{A}_{d,t}^{(\mathbf{k})}(W_i,x,\gamma^*(x))-\mathbb{E}[\widetilde{A}_{d,t}^{(\mathbf{k})}(W,x,\gamma^*(x))]\right|\\
&\le\max_{j\in\mathbb{N}_{J_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{A}_{d,t}^{(\mathbf{k})}(W_i,x_j,\gamma^*(x_j))-\mathbb{E}[\widetilde{A}_{d,t}^{(\mathbf{k})}(W,x_j,\gamma^*(x_j))]\right| &(3.8.41)\\
&\quad+\max_{j\in\mathbb{N}_{J_n}}\sup_{x\in\mathscr{B}(x_j,\delta_n)\cap\mathscr{X}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{A}_{d,t}^{(\mathbf{k})}(W_i,x,\gamma^*(x))-\widetilde{A}_{d,t}^{(\mathbf{k})}(W_i,x_j,\gamma^*(x_j))\right)\right| &(3.8.42)\\
&\quad+\max_{j\in\mathbb{N}_{J_n}}\sup_{x\in\mathscr{B}(x_j,\delta_n)\cap\mathscr{X}}\left|\mathbb{E}[\widetilde{A}_{d,t}^{(\mathbf{k})}(W,x,\gamma^*(x))]-\mathbb{E}[\widetilde{A}_{d,t}^{(\mathbf{k})}(W,x_j,\gamma^*(x_j))]\right|. &(3.8.43)
\end{aligned}$$

In view of (3.8.39), (3.8.42) is bounded from above by

$$\max_{i\in\mathbb{N}_n, j\in\mathbb{N}_{J_n}} \sup_{x\in\mathscr{B}\left(x_j,\delta_n\right)\cap\mathscr{X}} h^{-\upsilon_c}\left|\widetilde{A}^{\dagger,(\mathbf{k})}_{d,t}(W_i,x_c,\gamma^*(x))-\widetilde{A}^{\dagger,(\mathbf{k})}_{d,t}(W_i,x_{c,j},\gamma^*(x_j))\right| \leq \kappa_n\varepsilon/3.$$

Meanwhile, since $x_d = x_{d,j}$, whenever $x\in\mathscr{B}(x_j,\delta_n)$, (3.8.40) then implies that (3.8.43) $\leq \kappa_n\varepsilon/3$.

To bound (3.8.41), we apply Bernstein's inequality.[8] Since the support of K is bounded, we have that $|\widetilde{A}^{(\mathbf{k})}_{d,t}(W,x,\gamma^*(x))| \leq C\|K\|_\infty$, for a sufficiently large positive constant C. Additionally, standard calculation gives

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widetilde{A}_{d,t}(W,x,\gamma^*(x))\right] =&\mathbb{E}[(I_{d,t}-p(d,t,(X_c,x_d)))^2H(h)\underline{\mathbf{X}}(x_c)\underline{\mathbf{X}}(x_c)'H(h)K_h(\underline{\mathbf{X}}_i(x_c))^2\mathbb{1}\{X_d=x_d\}]\\ &+o\left(h^{-\upsilon_c}\right)\\ =&h^{-\upsilon_c}\mathscr{I}(x)_{\iota(d,t),\iota(d,t)}\mathbf{T}_p(x_c)f_X(x)+o\left(h^{-\upsilon_c}\right).\end{aligned}$$

Hence, $\mathbb{V}\text{ar}\left[\widetilde{A}^{(\mathbf{k})}_{d,t}(W,x,\gamma^*(x))\right] \leq Ch^{-\upsilon_c}$ under Assumption 3.5.4.

With these two results in hand, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{j\in\mathbb{N}_{J_n}}\left|\frac{1}{n}\sum_{i=1}^n\widetilde{A}^{(\mathbf{k})}_{d,t}(W_i,x_j,\gamma^*(x_j))-\mathbb{E}[\widetilde{A}^{(\mathbf{k})}_{d,t}(W,x_j,\gamma^*(x_j))]\right|\geq\kappa_n\varepsilon/3\right)\\ \leq&\sum_{j=1}^{J_n}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\widetilde{A}^{(\mathbf{k})}_{d,t}(W_i,x_j,\gamma^*(x_j))-\mathbb{E}[\widetilde{A}^{(\mathbf{k})}_{d,t}(W,x_j,\gamma^*(x_j))]\right|\geq\kappa_n\varepsilon/3\right)\\ \leq&2J_n\exp\left(-\frac{\varepsilon^2\log n}{C+C(\varepsilon\log n\cdot n^{-1}h^{-\upsilon_c})^{1/2}}\right)\leq\exp\left(-\frac{(\varepsilon^2-\kappa_b)\log n}{C}\right),\end{aligned}$$

where the first inequality is due to the Bonferoni inequality and the second is by Bernstein's inequality. The far right side goes to 0 when $\varepsilon^2>\kappa_b$. Hence, (3.8.41) $\leq \kappa_n\varepsilon/3$.

Combining (3.8.41)–(3.8.43) gives

$$\mathbb{P}\left(\sup_{x\in\mathscr{X}}\left|\frac{1}{n}\sum_{i=1}^n\widetilde{A}^{(\mathbf{k})}_{d,t}(W_i,x,\gamma^*(x))-\mathbb{E}[\widetilde{A}^{(\mathbf{k})}_{d,t}(W,x,\gamma^*(x))]\right|\geq\kappa_n\varepsilon\right)\to 0. \tag{3.8.44}$$

This complete the proof for (3.8.37).

Next, we establish (3.8.38). Define $I_o(x_d,z_d)=\sum_{s=1}^{\upsilon_o}\mathbb{1}\{|x_{o,s}-z_{o,s}|=1\}\prod_{l\neq s}\mathbb{1}\{x_{o,l}=z_{o,l}\}$, and $I_u(x_d,z_d)=$

[8] Let $\{X_i\}_{i=1}^n$ be independent zero-mean random variables. Suppose $|X_i|\leq M$ almost surely, for $i\in\mathbb{N}_n$. Then, Bernstein's inequality states that for all $t\geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i\geq t\right)\leq\exp\left(-\frac{t^2/2}{\sum_{i=1}^n\mathbb{E}[X_i^2]+Mt/3}\right).$$

$\sum_{s=1}^{v_u}\mathbb{1}\{x_{u,s}\neq z_{u,s}\}\prod_{l\neq s}\mathbb{1}\{x_{u,l}=z_{u,l}\}$. From a Taylor expansion of order $p+1$, we deduce that, uniformly in $x\in\mathscr{X}$,

$$\begin{aligned}
&\mathbb{E}[\widetilde{A}_{d,t}(W,x,\gamma^*(x))]\\
=&\frac{1}{(p+1)!}\sum_{(d',t')\in\mathscr{S}_-}\mathbb{E}\left[\mathscr{I}(X_c,x_d)_{\iota(d,t),\iota(d',t')}\mathbf{g}_{d',t'}^{(p+1)}(X_c,x_d)'\underline{\mathbf{X}}^{(p+1)}(x_c)H(h)\underline{\mathbf{X}}(x_c)K_h(\underline{\mathbf{X}}^{(1)}(x_c))\mathbb{1}\{X_d=x_d\}\right]\\
&+\sum_{z_d\in\mathscr{X}_d\backslash x_d}\sum_{j=o,u}\lambda_j I_j(x_d,z_d)\left(p(d,t,x)-p(d,t,(x_c,z_d))\right)\mathbb{E}\left[H(h)\underline{\mathbf{X}}(x_c)K_h(\underline{\mathbf{X}}^{(1)}(x_c))\mathbb{1}\{X_d=z_d\}\right]\\
&+(s.o.)\\
=&\frac{h^{p+1}}{(p+1)!}\sum_{(d',t')\in\mathscr{S}_-}\mathscr{I}(x)_{\iota(d,t),\iota(d',t')}\mathbf{M}_{p,p+1}(x_c)\mathbf{g}_{d',t'}^{(p+1)}(x)f_X(x)\\
&+\sum_{z_d\in\mathscr{X}_d\backslash x_d}\sum_{j=o,u}\lambda_j I_j(x_d,z_d)\left(p(d,t,x)-p(d,t,(x_c,z_d))\right)\mathbf{M}_{p,0}(x_c)f_X(x_c,z_d)\\
&+o(h^{p+1}+\lambda_o+\lambda_u)\\
=&O(h^{p+1}+\lambda_o+\lambda_u),
\end{aligned}$$

where $(s.o.)$ stands for smaller order terms. The last equality is due to Assumptions 3.5.2 and 3.5.4. ■

3.8.3 Auxiliary Lemmas and Results

3.8.3.1 Auxiliary Lemmas

Lemma 3.6 Under Assumptions 3.1 and 3.2, for $d,t\in\{0,1\}$ and any measurable function $h:\mathscr{X}\to\mathbb{R}$,

$$(i)\ \mathbb{E}\left[I_{d,t}(Y-m_{d,t}(X))h(X)\right]=0, \tag{3.8.45}$$

$$(ii)\ \mathbb{E}\left[\left(w_{1,1}-w_{d,t}\right)(W)h(X)\right]=0. \tag{3.8.46}$$

Proof of Lemma 3.6: This lemma follows immediately from the LIE. ■

Lemma 3.7 Suppose the conditions of Theorem 3.2 hold. Then, for $\widehat{w}$ defined in (3.3.1) with $\widehat{p}$ given by (3.3.7), we have

$$\mathbb{E}_n[(Y-m_{d,t}(X))\left(\widehat{w}_{d,t}-w_{d,t}\right)(W)]=o_p(n^{-1/2}),$$

for $(d,t)\in\mathscr{S}_-$.

Proof of Lemma 3.7:

Recall the definition of $w^{\dagger}$ as given in (3.8.15), and decompose the difference between $\widehat{w}_{d,t}$ and $w_{d,t}$ as

$$\begin{aligned}
&\mathbb{E}_n[(Y - m_{d,t}(X))\left(\widehat{w}_{d,t} - w_{d,t}\right)(W)] \\
&= \mathbb{E}_n\left[(Y - m_{d,t}(X))\left(w^{\dagger}_{d,t} - w_{d,t}\right)(W)\right] + \mathbb{E}_n\left[(Y - m_{d,t}(X))\left(\widehat{w}_{d,t} - w^{\dagger}_{d,t}\right)(W)\right] \\
&\equiv \Delta^1_w + \Delta^2_w.
\end{aligned}$$

We bound the two terms in turn. By a third-order Taylor expansion of Δ^1_w around $p(d,t,x)$, we get

$$\begin{aligned}
\Delta^1_w =& \mathbb{E}_n\left[\frac{I_{d,t}(Y - m_{d,t}(X))}{p(d,t,X)p(1,1)}\left(\widehat{p}(1,1,X) - p(1,1,X)\right)\right] \\
&- \mathbb{E}_n\left[\frac{I_{d,t}p(1,1,X)(Y - m_{d,t}(X))}{p^2(d,t,X)p(1,1)}\left(\widehat{p}(d,t,X) - p(d,t,X)\right)\right] + R_{n,d,t} \\
\equiv& \Delta^{11}_w + \Delta^{12}_w + R_{n,d,t},
\end{aligned}$$

where the remainder term, $R_{n,d,t}$, collects the second-order terms. Specifically,

$$\begin{aligned}
R_{n,d,t} =& \mathbb{E}_n\left[(Y - m_{d,t}(X))\frac{I_{d,t}}{p(1,1)}\left(-\frac{(\widehat{p}(1,1,X) - p(1,1,X))(\widehat{p}(d,t,X) - p(d,t,X))}{p^2(d,t,X)}\right)\right] \\
&+ \mathbb{E}_n\left[(Y - m_{d,t}(X))\frac{I_{d,t}}{p(1,1)}\left(\frac{p(1,1,X)(\widehat{p}(d,t,X) - p(d,t,X))^2}{\tilde{p}^3(d,t,X)}\right)\right],
\end{aligned}$$

where the intermediate point $\tilde{p}(d,t,x)$ lying between $\widehat{p}(d,t,x)$ and $p(d,t,x)$. Under Assumptions 3.2 (iii) and 3.5.1, both $\widehat{p}(d,t,x)$ and $p(d,t,x)$ are (asymptotically) bounded away from zero, uniformly over $\mathscr{X}$ and for $(d,t) \in \mathscr{S}$. Moreover, $\mathbb{E}[|Y - m_{d,t}(X)|] = O(1)$ under Assumption 3.5.3. We deduce that $R_{n,d,t} = O_p\left(\|\widehat{p}(1,1,\cdot) - p(1,1,\cdot)\|^2_{\infty}\right) + O_p\left(\|\widehat{p}(d,t,\cdot) - p(d,t,\cdot)\|^2_{\infty}\right)$, which is $o_p\left(n^{-1/2}\right)$ by Lemma 3.4 and Assumption 3.5.5.

The first two terms in the decomposition of Δ^1_w share a similar structure. We only derive the stochastic limit for Δ^{11}_w.

Using the asymptotic expansion of local polynomial estimators in Lemma 3.4, we obtain

$$\Delta^{11}_w = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I_{d,t,i}(Y_i - m_{d,t}(X_i))}{p(d,t,X_i)p(1,1)}\left(B^{(ps)}_{n,1,1}(X_i) + S^{(ps)}_{n,1,1}(X_i) + R^{(ps)}_{n,1,1}(X_i)\right)\right\}.$$

We proceed by establishing bounds for the convergence rate of the terms involving the bias, the first-order stochastic and the remainder, respectively.

To analyze the bias, we first apply Chebyshev's inequality and obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{I_{d,t,i}(Y_i-m_{d,t}(X_i))}{p(d,t,X_i)p(1,1)}B_{n,1,1}^{(ps)}(X_i) =& \mathbb{E}\left[\frac{I_{d,t}(Y-m_{d,t}(X))}{p(d,t,X)p(1,1)}B_{n,1,1}^{(ps)}(X)\right]\\ &+O_p\left(n^{-1/2}(h^{p+1}+\lambda_o+\lambda_u)\right),\end{aligned}$$

where the rate of the remainder comes from standard variance calculation. Owning to Lemma 3.6(i), the mean on the right-hand side is zero, which leads to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I_{d,t,i}(Y_i-m_{d,t}(X_i))}{p(d,t,X_i)p(1,1)}B_{n,1,1}^{(ps)}(X_i)=O_p\left(n^{-1/2}(h^{p+1}+\lambda_o+\lambda_u)\right). \tag{3.8.47}$$

Under the bandwidth restrictions in Assumption 3.5.5, this term is $o_p\left(n^{-1/2}\right)$.

We now introduce the term $\psi_{w1,d,t}(W_i,W_j)$, which represents the summand of the first-order stochastic term as follows

$$\psi_{w1,d,t}(W_i,W_j)=\frac{I_{d,t,i}(Y_i-m_{d,t}(X_i))}{p(d,t,X_i)p(1,1)}\left(G_{1,1}^{(ps)}(W_j,X_i)-\mathbb{E}[G_{1,1}^{(ps)}(W_j,X_i)|X_i]\right). \tag{3.8.48}$$

By its definition, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I_{d,t,i}(Y_i-m_{d,t}(X_i))}{p(d,t,X_i)p(1,1)}S_{n,1,1}^{(ps)}(X_i)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\psi_{w1,d,t}(W_i,W_j). \tag{3.8.49}$$

Given the construction, we have $\mathbb{E}\left[\psi_{w,d,t}(W_i,W_j)|W_i\right]=0$. Moreover, by Lemma 3.6 (i), we also have that $\mathbb{E}[\psi_{w,d,t}(W_i,W_j)|W_j]=0$. Hence, (3.8.49) represents a second-order U-statistic with first-order degenerate kernel. Lemma 3.3 and standard variance calculation then gives that

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\psi_{w1,d,t}(W_i,W_j)=O_p\left(n^{-1}h^{-\upsilon_c/2}\right), \tag{3.8.50}$$

Under our bandwidth assumptions, this term is $o_p\left(n^{-1/2}\right)$.

Under Assumption 3.2(iii), $p(d,t,x)$ is uniformly bounded away from zero for all $x\in\mathscr{X}$ and for all $(d,t)\in\mathscr{S}_{-}$. Also, under Assumption 3.5.3, we have $\mathbb{E}[|Y-m_{d,t}(X)|]=O(1)$. Consequently, we can deduce that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{I_{d,t,i}(Y_i-m_{d,t}(X_i))}{p(d,t,X_i)p(1,1)}R_{n,1,1}^{(ps)}(X_i) &= O_p\left(\sup_{i\in\mathbb{N}_n}\left|R_{n,1,1}^{(ps)}(X_i)\right|\right)\\ &= O_p\left(\left(h^{p+1}+\lambda_o+\lambda_u+\sqrt{\log n/(nh^{\upsilon_c})}\right)^2\right)\end{aligned} \tag{3.8.51}$$

which is $o_p\left(n^{-1/2}\right)$ under Assumption 3.5.5.

Combining (3.8.47), (3.8.50), and (3.8.51), we can conclude that $\Delta_w^{11} = o_p\left(n^{-1/2}\right)$.

By the same reasoning, we can demonstrate that Δ_w^{12} is dominated by the first-order stochastic term. Define

$$\psi_{w2,d,t}(W_i,W_j) = -\frac{I_{d,t}p(1,1,X_i)(Y_i - m_{d,t}(X_i))}{p^2(d,t,X_i)p(1,1)}\left(G_{d,t}^{(ps)}(W_j,X_i) - \mathbb{E}[G_{d,t}^{(ps)}(W_j,X_i)|X_i]\right), \tag{3.8.52}$$

As a result, the leading term is given by $n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\psi_{w2,d,t}(W_i,W_j)$, which has an order of $O_p\left(n^{-1}h^{-\upsilon_c/2}\right)$ $= o_p\left(n^{-1/2}\right)$. The detailed proof is omitted for brevity.

Now, let's consider Δ_w^2. Define $\widehat{p}(1,1) = \mathbb{E}_n\left[\dfrac{I_{d,t}\widehat{p}(1,1,X)}{\widehat{p}(d,t,X)}\right]$.

$$\begin{aligned}\Delta_w^2 &= \mathbb{E}_n\left[\frac{I_{d,t}\widehat{p}(1,1,X)(Y - m_{d,t}(X))}{\widehat{p}(d,t,X)}\left(\frac{1}{\widehat{p}(1,1)} - \frac{1}{p(1,1)}\right)\right]\\ &= \mathbb{E}_n\left[\frac{I_{d,t}\widehat{p}(1,1,X)(Y - m_{d,t}(X))}{\widehat{p}(d,t,X)}\right]\cdot O_p\left(|\widehat{p}(1,1) - p(1,1)|\right),\end{aligned}$$

where the second line follows by a first-order Taylor expansion of the right-hand side of the first equality in $\widehat{p}(1,1)$ around $p(1,1)$. In the proof of Lemma 3.1, it is established that when $\widehat{p}$ is uniformly convergent to p, $|\widehat{p}(1,1) - p(1,1)| = o_p(1)$. The uniform convergence follows by Lemma 3.4 under the rate conditions specified in Assumption 3.5.5.

To study the first term, we can use an approach similar to the proof of Δ_w^1, and show that

$$\mathbb{E}_n\left[\frac{I_{d,t}\widehat{p}(1,1,X)(Y - m_{d,t}(X))}{\widehat{p}(d,t,X)}\right] = \mathbb{E}_n\left[\frac{I_{d,t}p(1,1,X)(Y - m_{d,t}(X))}{p(d,t,X)}\right] + o_p\left(n^{-1/2}\right).$$

Due to Lemma 3.6(i), the first term on the right-hand side of the preceding equation has a mean of zero. Consequently, this term is of order $O_p\left(n^{-1/2}\right)$. This completes our proof. ■

Lemma 3.8 Suppose the conditions of Theorem 3.2 hold, then with $\widehat{m}$ given by (3.3.9),

$$\mathbb{E}_n[(w_{1,1} - w_{d,t})(W)\cdot\left(\widehat{m}_{d,t} - m_{d,t}\right)] = o_p(n^{-1/2}),$$

for $(d,t) \in \mathscr{S}_{-}$.

Proof of Lemma 3.8:

The proof closely resembles the first part of Lemma 3.7. We first decompose the estimation error for the OR functions as

$$\mathbb{E}_n[(w_{1,1}-w_{d,t})(W)\left(\widehat{m}_{d,t}-m_{d,t}\right)(X)] = \frac{1}{n}\sum_{i=1}^{n}\left\{(w_{1,1}-w_{d,t})(W_i)\left(B_{n,d,t}^{(or)}(X_i)+S_{n,d,t}^{(or)}(X_i)+R_{n,d,t}^{(or)}(X_i)\right)\right\}.$$

We address the three terms individually. For the bias term

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left\{(w_{1,1}-w_{d,t})(W_i)B_{n,d,t}^{(or)}(X_i)\right\} &= \mathbb{E}\left[(w_{1,1}-w_{d,t})(W)B_{n,d,t}^{(or)}(X)\right]+O_p\left(n^{-1/2}(b_{d,t}^{q+1}+\vartheta_{o,d,t}+\vartheta_{u,d,t})\right)\\ &= O_p\left(n^{-1/2}(b_{d,t}^{q+1}+\vartheta_{o,d,t}+\vartheta_{u,d,t})\right)=o_p\left(n^{-1/2}\right),\end{aligned}$$

where the first equality follows from Chebyshev's inequality, and the second is derived from Lemma 3.6(ii).

Next, for the first-order stochastic term, we define

$$\psi_{m,d,t}(W_i,W_j) = (w_{1,1}-w_{d,t})(W_i)\left(G_{d,t}^{(or)}(W_j,X_i)-\mathbb{E}[G_{d,t}^{(or)}(W_j,X_i)|X_i]\right), \tag{3.8.53}$$

By definition,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{(w_{1,1}-w_{d,t})(W_i)S_{n,d,t}^{(or)}(X_i)\right\} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\psi_{m,d,t}(W_i,W_j).$$

In view of Lemma 3.6(ii), the right-hand side of the above equation is a second-order U-statistic with a degenerate first-order kernel. A standard variance calculation shows that it is of the order $O_p\left(n^{-1}b_{d,t}^{\upsilon_c/2}\right)$, which is $o_p\left(n^{-1/2}\right)$ due to our bandwidth restrictions.

Finally, as $p(d,t,x)$ is uniformly bounded away from zero under Assumption 3.2(iii), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left\{(w_{1,1}-w_{d,t})(W_i)R_{n,d,t}^{(or)}(X_i)\right\} &= O_p\left(\sup_{i\in\mathbb{N}_n}\left|R_{n,d,t}^{(or)}(X_i)\right|\right)\\ &= O_p\left(\left((b_{d,t}^{q+1}+\vartheta_{o,d,t}+\vartheta_{u,d,t})+\sqrt{\log n/\left(nb_{d,t}^{\upsilon_c}\right)}\right)^2\right),\end{aligned}$$

which is $o_p\left(n^{-1/2}\right)$ under Assumption 3.5.5. This completes our proof. ■

3.8.3.2 Mean Integrated Squared Error

Cross-validated bandwidth asymptotically minimizes the mean integrated squared errors (MISE). Given user-specified weight functions $\omega^{ps}(\cdot), \omega^{or}_{d,t}(\cdot) : \mathscr{X} \to \mathbb{R}_+$, MISE is defined as

$$\begin{aligned}\chi(h, \lambda, \{b_{d,t}, \vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-}) = &\int_{\mathscr{X}} \mathbb{E}\left[\|\hat{\mathbf{p}}_-(x) - \mathbf{p}_-(x)\|^2\right] \omega^{ps}(x)dx \\ &+ \sum_{(d,t)\in\mathscr{S}_-} \int_{\mathscr{X}} \mathbb{E}\left[\left|\hat{m}_{d,t}(x) - m_{d,t}(x)\right|^2\right] \omega^{or}_{d,t}(x)dx.\end{aligned}$$

Let $(h^*, \lambda^*, \{b^*_{d,t}, \vartheta^*_{d,t}\}_{(d,t)\in\mathscr{S}_-})$ denote the minimizer of the MISE. In the subsequent analysis, we investigate the properties of these optimal smoothing parameters.

For $(d,t) \in \mathscr{S}_-$, we represent the $n_k \times 1$ vector of k-th derivatives $p(d,t,x)$ as $\mathbf{p}^{(k)}_{d,t}(x)$, ordered lexicographically according to the method discussed earlier in the paper. Define $\mathbf{g}^{(k)}_-(x) = \left(\mathbf{g}^{(k)}_{1,0}(x), \mathbf{g}^{(k)}_{0,1}(x), \mathbf{g}^{(k)}_{0,0}(x)\right)$. For $j = p, q$, let $\rho^b_{j,1}(x_c) = e'_{N_j,1}\mathbf{Q}_j(x_c)^{-1}\mathbf{M}_{j,j+1}(x_c)$, $\rho^b_{j,2}(x_c) = e'_{N_j,1}\mathbf{Q}_j(x_c)^{-1}\mathbf{M}_{j,0}(x_c)$, and $\rho^v_j(x_c) = e'_{N_j,1}\mathbf{Q}_j(x_c)^{-1}\mathbf{T}_j(x_c)\mathbf{Q}_j(x_c)^{-1}\cdot e_{N_j,1}$. Additionally, we define terms associated with the asymptotic bias and variance of $\hat{\mathbf{p}}_-(x) - \mathbf{p}_-(x)$ as follows

$$\begin{aligned}\mathscr{B}^{ps}(x,h,\lambda) =& \frac{h^{p+1}}{(p+1)!}\rho^b_{p,1}(x_c)\mathbf{g}^{(p+1)}_-(x)\mathscr{I}(x) \\ &+ \sum_{z_d\in\mathscr{X}_d\setminus x_d}\sum_{j=o,u}\frac{f_X(x_c,z_d)}{f_X(x)}\lambda_j I_j(x_d,z_d)\rho^b_{p,2}(x_c)\left(\mathbf{p}_-(x) - \mathbf{p}_-(x_c,z_d)\right), \\ \mathscr{V}^{ps}(x,h,\lambda) =& \frac{\mathscr{I}(x)\rho^v_p(x_c)}{h^{\upsilon_c}f_X(x)}.\end{aligned}$$

For the OR functions, we define

$$\begin{aligned}\mathscr{B}^{or}_{d,t}(x,b,\vartheta) =& \frac{b^{q+1}}{(q+1)!}\left(\rho^b_{q,1}(x_c)\mathbf{m}^{(q+1)}_{d,t}(x)\right) \\ &+ \sum_{z_d\in\mathscr{X}_d\setminus x_d}\sum_{j=o,u}\frac{f_X(x_c,z_d)}{f_X(x)}\vartheta_j I_j(x_d,z_d)\rho^b_{q,2}(x_c)\left(m_{d,t}(x) - m_{d,t}(x_c,z_d)\right), \\ \mathscr{V}^{or}_{d,t}(x,b,\vartheta) =& \frac{\sigma^2_{d,t}(x)\rho^v_q(x_c)}{b^{\upsilon_c}f_X(x)},\end{aligned}$$

where $\sigma^2_{d,t}(x) = \mathbb{E}[I_{d,t}(Y - m_{d,t}(X))^2|X = x]$.

Finally, we define a first-order approximation of the MISE as

$$\begin{aligned}\chi^*(h,\lambda,\{b_{d,t},\vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-}) = &\int_{\mathscr{X}}\left\{\|\mathscr{B}^{ps}(x,h,\lambda)\|^2 + \mathrm{tr}(\mathscr{V}^{ps}(x,h,\lambda))\right\}\omega^{ps}(x)dx \\ &+ \sum_{(d,t)\in\mathscr{S}_-}\int_{\mathscr{X}}\left\{\mathscr{B}^{or}_{d,t}(x,b_{d,t},\vartheta_{d,t})^2 + \mathscr{V}^{or}_{d,t}(x,b_{d,t},\vartheta_{d,t})\right\}\omega^{or}_{d,t}(x)dx. \qquad (3.8.54)\end{aligned}$$

We denote the constrained minimizer of χ^* as $(h^o, \lambda^o, b^o_{d,t}, \vartheta^o_{d,t}{}_{(d,t)\in\mathscr{S}_-})$, where each argument of the function is constrained to be non-negative.

Assumption 3.6
1. The constrained minimizer of χ^*, denoted as $(h^o, \lambda^o, \{b^o_{d,t}, \vartheta^o_{d,t}\}_{(d,t)\in\mathscr{S}_-})$, is uniquely determined and finite.
2. The constrained minimizer resides in $[0, \delta_n]^{12}$, where $n^{\varepsilon}\delta_n \to \infty$ for any $\varepsilon > 0$.

Theorem 3.4 Assuming that Assumptions 3.1, 3.5, and 3.6 hold and both p and q are odd, the optimal bandwidths $(h^*, \lambda^*, \{b^*_{d,t}, \vartheta^*_{d,t}\}_{(d,t)\in\mathscr{S}_-})$ satisfy

$$\begin{aligned} h^* &\sim h^o n^{-1/(2p+\upsilon_c+2)}, & \lambda^* &\sim \lambda^o n^{-2/(2p+\upsilon_c+2)}, \\ b^*_{d,t} &\sim b^o_{d,t} n^{-1/(2q+\upsilon_c+2)}, & \vartheta^*_{d,t} &\sim \vartheta^o_{d,t} n^{-2/(2q+\upsilon_c+2)}, \text{ for } (d,t) \in \mathscr{S}_-. \end{aligned}$$

Proof of Theorem 3.4:

From the uniform linear expansions of Lemma 3.4, we know that

$$\mathbb{E}\left[\|\hat{\mathbf{p}}_-(x) - \mathbf{p}_-(x)\|^2\right] = \|\mathbb{E}[\mathscr{I}(x)\mathbf{A}_-(W,x)]\|^2 + n^{-1}\operatorname{tr}(\mathbb{V}\mathrm{ar}[\mathscr{I}(x)\mathbf{A}_-(W,x)]) + (s.o.),$$

where

$$\begin{aligned} &\mathbb{E}[\mathscr{I}(x)\mathbf{A}_-(W,x)] = \mathscr{I}(x)(I_3 \otimes e_{N_p,1})'\Sigma^{ps}(x)^{-1}\,\mathbb{E}[\widetilde{\boldsymbol{A}}_-(W,x,\gamma^*(x))] \\ =&\frac{h^{p+1}}{(p+1)!}\mathscr{I}(x)(I_3 \otimes e_{N_p,1})'(\mathscr{I}(x)\otimes\mathbf{Q}_p(x_c)f_X(x))^{-1}\Big\{(\mathscr{I}(x)\otimes\mathbf{M}_{p,p+1}(x_c))\operatorname{vec}\left(\mathbf{g}_-^{(p+1)}(x)\right)f_X(x) \\ &+\sum_{z_d\in\mathscr{X}_d\setminus x_d}\sum_{j=o,u}\lambda_j I_j(x_d,z_d)\left(\mathbf{p}_-(x)-\mathbf{p}_-(x_c,z_d)\right)\otimes\mathbf{M}_{p,0}(x_c)f_X(x_c,z_d)\Big\} + o\left(h^{p+1}+\lambda_o+\lambda_u\right) \\ =&\frac{h^{p+1}}{(p+1)!}e'_{N_p,1}\mathbf{Q}_p(x_c)^{-1}\mathbf{M}_{p,p+1}(x_c)\mathbf{g}_-^{(p+1)}(x)\mathscr{I}(x) \\ &+\sum_{z_d\in\mathscr{X}_d\setminus x_d}\sum_{j=o,u}\frac{f_X(x_c,z_d)}{f_X(x)}\lambda_j I_j(x_d,z_d)e'_{N_p,1}\mathbf{Q}_p(x_c)^{-1}\mathbf{M}_{p,0}(x_c)\left(\mathbf{p}_-(x)-\mathbf{p}_-(x_c,z_d)\right) \\ &+o\left(h^{p+1}+\lambda_o+\lambda_u\right) \\ =&\mathscr{B}^{ps}(x,h,\lambda)+o\left(h^{p+1}+\lambda_o+\lambda_u\right), \end{aligned} \tag{3.8.55}$$

and

$$\mathbb{V}\mathrm{ar}[\mathscr{I}(x)\mathbf{A}_-(W,x)] = h^{-\upsilon_c}\mathscr{I}(x)(I_3\otimes e_{N_p,1})'\Sigma^{ps}(x)^{-1}\left(\mathscr{I}(x)\otimes\mathbf{T}_p(x_c)f_X(x)\right)\Sigma^{ps}(x)^{-1}(I_3\otimes e_{N_p,1})\mathscr{I}(x)$$

$$
\begin{aligned}
&=h^{-\upsilon_c}\mathscr{I}(x)(I_3\otimes e_{N_p,1})'(\mathscr{I}(x)\otimes\mathbf{Q}_p(x_c)f_X(x))^{-1}\left(\mathscr{I}(x)\otimes\mathbf{T}_p(x_c)f_X(x)\right)\\
&\quad\cdot\left(\mathscr{I}(x)\otimes\mathbf{Q}_p(x_c)f_X(x)\right)^{-1}(I_3\otimes e_{N_p,1})\mathscr{I}(x)+o\left(h^{-\upsilon_c}\right)\\
&=h^{-\upsilon_c}f_X(x)^{-1}\mathscr{I}(x)e'_{N_p,1}\mathbf{Q}_p(x_c)^{-1}\mathbf{T}_p(x_c)\mathbf{Q}_p(x_c)^{-1}e_{N_p,1}+o\left(h^{-\upsilon_c}\right)\\
&=\mathscr{V}^{ps}(x,h,\lambda)+o\left(h^{-\upsilon_c}\right). \qquad (3.8.56)
\end{aligned}
$$

Analogously, for $(d,t)\in\mathscr{S}_-$

$$
\mathbb{E}\left[\left|\widehat{m}_{d,t}(x)-m_{d,t}(x)\right|^2\right]=\left|\mathbb{E}[G^{(or)}_{d,t}(W,x)]\right|^2+n^{-1}\mathbb{V}\text{ar}\left[G^{(or)}_{d,t}(W,x)\right]+(s.o.),
$$

where

$$
\begin{aligned}
&\mathbb{E}[G^{(or)}_{d,t}(W,x)]=e'_{N_q,1}\Sigma^{or}_{d,t}(x)^{-1}\,\mathbb{E}[H(b_{d,t})\underline{\mathbf{X}}(X_j)I_{d,t}\xi^{or}_{d,t}(x)\widetilde{K}_{or}(X;x,b_{d,t},\vartheta_{d,t})]\\
&=\frac{b^{q+1}_{d,t}}{(q+1)!}e'_{N_q,1}(\mathbf{Q}_q(x_c)f_X(x))^{-1}\left\{\mathbf{M}_{q,q+1}(x_c)\mathbf{m}^{(q+1)}_{d,t}(x)f_X(x)\right.\\
&\quad\left.+\sum_{z_d\in\mathscr{X}_d\backslash x_d}\sum_{j=o,u}\vartheta_{d,t,j}I_j(x_d,z_d)\left(m_{d,t}(x)-m_{d,t}(x_c,z_d)\right)\mathbf{M}_{q,0}(x_c)f_X(x_c,z_d)\right\}\\
&\quad+o\left(b^{q+1}_{d,t}+\vartheta_{d,t,o}+\vartheta_{d,t,u}\right)\\
&=\frac{b^{q+1}_{d,t}}{(q+1)!}\left(e'_{N_q,1}\mathbf{Q}_q(x_c)^{-1}\mathbf{M}_{q,q+1}(x_c)\mathbf{m}^{(q+1)}_{d,t}(x)\right)\\
&\quad+\sum_{z_d\in\mathscr{X}_d\backslash x_d}\sum_{j=o,u}\frac{f_X(x_c,z_d)}{f_X(x)}\vartheta_jI_j(x_d,z_d)e'_{N_q,1}\mathbf{Q}_q(x_c)^{-1}\mathbf{M}_{q,0}(x_c)\left(m_{d,t}(x)-m_{d,t}(x_c,z_d)\right)\\
&\quad+o\left(b^{q+1}_{d,t}+\vartheta_{d,t,o}+\vartheta_{d,t,u}\right),\\
&=\mathscr{B}^{or}_{d,t}(x,b_{d,t},\vartheta_{d,t})+o\left(b^{q+1}_{d,t}+\vartheta_{d,t,o}+\vartheta_{d,t,u}\right), \qquad (3.8.57)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{V}\text{ar}\left[G^{(or)}_{d,t}(W,x)\right]&=b^{-\upsilon_c}_{d,t}e'_{N_q,1}\Sigma^{or}_{d,t}(x)^{-1}\,\mathbb{E}[H(b_{d,t})\underline{\mathbf{X}}(X_j)I_{d,t}(Y-m_{d,t}(X))^2\\
&\quad+H(b_{d,t})\underline{\mathbf{X}}(X_j)'\widetilde{K}_{or}(X;x,b_{d,t},\vartheta_{d,t})^2]\Sigma^{or}_{d,t}(x)^{-1}e_{N_q,1}+o\left(b^{-\upsilon_c}\right)\\
&=b^{-\upsilon_c}_{d,t}e'_{N_q,1}(\mathbf{Q}_q(x_c)f_X(x))^{-1}\left(\sigma^2_{d,t}(x)\mathbf{T}_q(x_c)f_X(x)\right)(\mathbf{Q}_q(x_c)f_X(x))^{-1}+o\left(b^{-\upsilon_c}\right)\\
&=b^{-\upsilon_c}_{d,t}f_X(x)^{-1}\sigma^2_{d,t}(x)e'_{N_q,1}\mathbf{Q}_q(x_c)^{-1}\mathbf{T}_q(x_c)\mathbf{Q}_q(x_c)^{-1}e_{N_q,1}+o\left(b^{-\upsilon_c}\right)\\
&=\mathscr{V}^{or}_{d,t}(x,b_{d,t},\vartheta_{d,t})+o\left(b^{-\upsilon_c}\right). \qquad (3.8.58)
\end{aligned}
$$

Now, we define

$$(h^{\dagger},\lambda^{\dagger},\{b^{\dagger}_{d,t},\vartheta^{\dagger}_{d,t}\}_{(d,t)\in\mathscr{S}_-}) = (n^{1/(2p+\upsilon_c+2)}h, n^{2/(2p+\upsilon_c+2)}\lambda, \{n^{1/(2q+\upsilon_c+2)}b_{d,t}, n^{2/(2q+\upsilon_c+2)}\vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-}).$$

It follows from (3.8.55)–(3.8.58) and standard analysis that

$$\begin{aligned}
\chi(h,\lambda,&\{b_{d,t},\vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-}) \\
=&n^{-2(p+1)/(2p+\upsilon_c+2)}\int_{\mathscr{X}}\left\{\left\|\mathscr{B}^{ps}(x,h^{\dagger},\lambda^{\dagger})\right\|^2+\operatorname{tr}(\mathscr{V}^{ps}(x,h^{\dagger},\lambda^{\dagger}))\right\}\omega^{ps}(x)dx \\
&+o\left(h^{p+1}+\lambda_o+\lambda_u+h^{-\upsilon_c}\right) \\
&+n^{-2(q+1)/(2q+\upsilon_c+2)}\sum_{(d,t)\in\mathscr{S}_-}\int_{\mathscr{X}}\left\{\mathscr{B}^{or}_{d,t}(x,b^{\dagger}_{d,t},\vartheta^{\dagger}_{d,t})^2+\mathscr{V}^{or}_{d,t}(x,b^{\dagger}_{d,t},\vartheta^{\dagger}_{d,t})\right\}\omega^{or}_{d,t}(x)dx \\
&+o\left(\sum_{(d,t)\in\mathscr{S}_-}\left\{b^{q+1}_{d,t}+\vartheta_{d,t,o}+\vartheta_{d,t,u}+b^{-\upsilon_c}_{d,t}\right\}\right),
\end{aligned}$$

uniformly over $[0,\delta_n]^{12}$. Since χ^* is separable in (h,λ) and $(\{b_{d,t},\vartheta_{d,t}\}_{(d,t)\in\mathscr{S}_-})$, and its constrained minimizer is well-defined, unique, and finite under Assumption 3.6, the proof is completed by minimizing χ with respect to $(h^{\dagger},\lambda^{\dagger},\{b^{\dagger}_{d,t},\vartheta^{\dagger}_{d,t}\}_{(d,t)\in\mathscr{S}_-})$ and recalling the definition of $(h^{o},\lambda^{o},\{b^{o}_{d,t},\vartheta^{o}_{d,t}\}_{(d,t)\in\mathscr{S}_-})$. ■

3.8.3.3 Plug-In Estimators

When employing the frequency method (i.e., $\lambda=\vartheta_{d,t}=0$), a straightforward plug-in rule can be used to determine the bandwidths $(h,\{b_{d,t}\}_{(d,t)\in\mathscr{S}_-})$. Notably, local polynomial estimators with an odd degree of fit are adaptive to boundaries, implying that the convergence rate of bias and variance remains constant regardless of the location of x. By solving Equation (3.8.54) and applying Theorem 3.4, the following results are obtained

$$\begin{aligned}
h^* &= \left(\frac{\int\left\|\rho^{b}_{p,1}(x_c)\mathbf{g}^{(p+1)}_{-}(x)\mathscr{I}(x)\right\|^2\omega^{ps}(x)dx}{\int\operatorname{tr}\left(\mathscr{I}(x)\rho^{v}_{p}(x_c)\right)/f_X(x)\cdot\omega^{ps}(x)dx}\frac{2(p+1)n}{\upsilon_c\{(p+1)!\}^2}\right)^{-1/(2p+\upsilon_c+2)}, \\
b^*_{d,t} &= \left(\frac{\int\left\|\rho^{b}_{q,1}(x_c)\mathbf{m}^{(q+1)}_{d,t}(x)\right\|^2\omega^{or}_{d,t}(x)dx}{\int\rho^{v}_{q}(x_c)/f_X(x)\cdot\omega^{or}_{d,t}(x)dx}\frac{2(q+1)n}{\upsilon_c\{(q+1)!\}^2}\right)^{-1/(2q+\upsilon_c+2)}, \text{ for } (d,t)\in\mathscr{S}_-.
\end{aligned}$$

These bandwidths, however, are infeasible due to the presence of unknown quantities related to the derivatives of the nuisance functions and local Fisher information. To estimate the optimal bandwidths, preliminary approximations of these quantities are necessary. An additional challenge arises from the complicated dependence of the plug-in bandwidths on the location of x (through ρ^b and ρ^v). One possible solution is to substitute the values evaluated at a boundary point with those associated with interior points. This replacement has a negligible impact on the consistency of the optimal bandwidth in general. The bandwidth selection process can be outlined in the following algorithm:

Algorithm 3.8.1 1. Let $\mathscr{X}_o$ collect all the unique values of $\{X_i\}_{i=1}^n$. Construct standard kernel estimates of covariate density with mixed data, $\widehat{f}_X(x)$, for $x \in \mathscr{X}_o$, following, e.g., Racine and Li (2004).

2. Use a polynomial multinomial logit regression of order $\ell = p+2$ to get preliminary estimates $\breve{\mathscr{I}}(x), \breve{\mathbf{g}}_-^{(p+1)}(x)$, $\breve{\mathbf{g}}_-^{(p+2)}(x)$, for $x \in \mathscr{X}_o$. Run polynomial regressions of order $\ell = q+2$ to obtain $\breve{\mathbf{m}}_{d,t}^{(q+1)}(x)$ and $\breve{\mathbf{m}}_{d,t}^{(q+2)}(x)$, for $x \in \mathscr{X}_o$.

3. Compute preliminary bandwidths

$$\breve{h} = \left(\frac{\mathbb{E}_n \left[\left\| \rho_{p,1}^b \breve{\mathbf{g}}_-^{(p+1)}(X) \breve{\mathscr{I}}(X) \right\|^2 \right]}{\rho_p^v \mathbb{E}_n \left[\widehat{f}_X^{-1}(X) \operatorname{tr} \left(\breve{\mathscr{I}}(X) \right) \right]} \frac{2(p+1)n}{\upsilon \{(p+1)!\}^2} \right)^{-1/(2p+\upsilon+2)},$$

$$\breve{b}_{d,t} = \left(\frac{\mathbb{E}_n \left[\left\| \rho_{q,1}^b \breve{\mathbf{m}}_{d,t}^{(q+1)}(X) \right\|^2 \right]}{\rho_q^v \mathbb{E}_n \left[\widehat{f}_X^{-1}(X) \right]} \frac{2(q+1)n}{\upsilon \{(q+1)!\}^2} \right)^{-1/(2q+\upsilon+2)},$$

$$\tilde{h} = \left(\frac{\mathbb{E}_n \left[\left\| \boldsymbol{\rho}_{p+1}^b \breve{\mathbf{g}}_-^{(p+2)}(X) \right\|^2 \right]}{\mathbb{E}_n \left[\widehat{f}_X^{-1}(X) \operatorname{tr} \left(\breve{\mathscr{I}}(X)^{-1} \otimes \boldsymbol{\rho}_{p+1}^v \right) \right]} \frac{2n}{\upsilon(2p+3)[(p+2)!]} \right)^{-1/(2p+\upsilon+4)},$$

$$\tilde{b}_{d,t} = \left(\frac{\mathbb{E}_n \left[\left\| \boldsymbol{\rho}_{q+1}^b \breve{\mathbf{m}}_{d,t}^{(q+2)}(X) \right\|^2 \right]}{\mathbb{E}_n \left[\left\| \widehat{f}_X^{-1}(X) \boldsymbol{\rho}_{q+1}^v \right\| \right]} \frac{2n}{\upsilon(2q+3)[(q+2)!]} \right)^{-1/(2q+\upsilon+4)},$$

where we omitted the dependence of ρ^b and ρ^v on x_c to signify that the boundary effect is disregarded. Furthermore, in the preceding equations, $\boldsymbol{\rho}_j^b = I'_{N_j,\mathbf{j}} \mathbf{Q}_j^{-1} \mathbf{M}_{j,j+1}$, $\boldsymbol{\rho}_j^v = I'_{N_j,\mathbf{j}} \mathbf{Q}_j^{-1} \mathbf{T}_j \mathbf{Q}_j^{-1} I_{N_j,\mathbf{j}}$, and $I_{N_j,\mathbf{j}}$ is a $N_j \times n_j$ matrix consisting of the last n_j columns of the $N_j \times N_j$ identity matrix.

4. Run a local polynomial logistic regression of order $\ell = p+1$, with bandwidth $\tilde{h}$, to obtain $\widehat{\mathbf{g}}_-^{(p+1)}(x)$. For each $(d,t) \in \mathscr{S}_-$, run a local polynomial regression of order $\ell = q+1$, using bandwidth $\widehat{b}_{d,t}$, to get $\widehat{\mathbf{m}}_{d,t}^{(q+1)}(x)$, for $x \in \mathscr{X}_o$.

5. Run a local polynomial logistic regression of order $\ell = p$, with bandwidth $\breve{h}$, to obtain $\widehat{\mathscr{I}}(x)$, for $x \in \mathscr{X}_o$.

6. Compute the optimal bandwidth $\widehat{h}$ and $\widehat{b}_{d,t}$, following

$$\widehat{h} = \left(\frac{\mathbb{E}_n \left[\left\| \rho^b_{p,1} \widehat{\mathbf{g}}_{-}^{(p+1)}(X) \widehat{\mathscr{I}}(X) \right\|^2 \right]}{\rho^v_p \mathbb{E}_n \left[\widehat{f}_X^{-1}(X) \operatorname{tr}\left(\widehat{\mathscr{I}}(X) \right) \right]} \frac{2(p+1)n}{\upsilon \{(p+1)!\}^2} \right)^{-1/(2p+\upsilon+2)} ,$$

$$\widehat{b}_{d,t} = \left(\frac{\mathbb{E}_n \left[\left\| \rho^b_{q,1} \widehat{\mathbf{m}}_{d,t}^{(q+1)}(X) \right\|^2 \right]}{\rho^v_q \mathbb{E}_n \left[\widehat{f}_X^{-1}(X) \right]} \frac{2(q+1)n}{\upsilon \{(q+1)!\}^2} \right)^{-1/(2q+\upsilon+2)} .$$

3.8.3.4 Cluster-Robust Inference: Bootstrap Procedures

In this section, we introduce two bootstrap procedures that are suitable for cluster-robust inference. The first algorithm uses a multiplier-bootstrap method to compute studentized and cluster-robust standard errors. This method has been previously described in Kline and Santos (2012) and Callaway et al. (2018). The second procedure is a bootstrap Hausman-type test, which provides bootstrapped p-values.

Let $V_{i=1}^n$ be a sequence of $i.i.d.$ random variables with zero mean and unit variance, which is independent of the original sample. One example is $i.i.d.$ Bernoulli random variables with $P(V = v_0) = 1 - v_0/\sqrt{5}$ and $P(V = 1 - v_0) = v_0/\sqrt{5}$, where $v_0 = (\sqrt{5}+1)/2$, as suggested by Mammen (1993). Now, given a generic ATT estimator, $\widehat{\tau}$, and an estimator of its influence function, $\widehat{\eta}(\cdot)$, we compute the clustered standard errors as follows:

Algorithm 3.8.2 1. In iteration b, draw a realization of V_b for each cluster. All observations within the same cluster share the same value of V_b.

2. Calculate a bootstrap estimate for ATT as

$$\widehat{\tau}_b^* = \widehat{\tau} + \mathbb{E}_n[V_b \cdot \widehat{\eta}(W)].$$

Form a bootstrap draw of the limiting distribution as

$$\widehat{R}_b^* = \sqrt{n}\,(\widehat{\tau}_b^* - \widehat{\tau}) .$$

3. Repeat Steps 1-2 B times.

4. Calculate the bootstrapped standard error, $\hat{\sigma}^*$, as the bootstrap interquartile range normalized by the interquartile range of the standard normal distribution: $\hat{\sigma}^* = (q_{0.75}(\widehat{R}) - q_{0.25}(\widehat{R}))/(z_{0.75} - z_{0.25})$, where $q_p(\widehat{R})$ is the p-th

sample quantile of the $\widehat{R}_b$ in the B draws, and z_p is the p-th quantile of the standard normal distribution.

Given the two DR DID estimators, $\widehat{\tau}_{dr}$ based on (3.3.1), $\widehat{\tau}_{sz}$ based on (3.4.1), and their respective linear expansions, $\widehat{\eta}_{dr}(\cdot)$ given in (3.3.11) and $\widehat{\eta}_{sz}(\cdot)$ given in (3.4.3), we conduct a cluster-robust Hausman-type test as follows

Algorithm 3.8.3

1. Calculate the Hausman test statistic, $\mathscr{T}_n$, following (3.4.2).

2. In iteration b, generate a realization of V_b for each cluster. Observations within the same cluster share the same value of V_b.

3. Calculate bootstrap estimates of the ATT as

$$\begin{aligned}\widehat{\tau}^*_{j,b} &= \widehat{\tau}_j + \mathbb{E}_n[V_b \cdot \widehat{\eta}_j(W)],\\ \hat{V}^*_b &= \mathbb{E}_n[V_b \cdot (\widehat{\eta}_{eff}(W) - \widehat{\eta}_{sz}(W))^2].\end{aligned}$$

   Form a bootstrap test statistic, $\mathscr{T}^*_b$, as

$$\mathscr{T}^*_b = n\left(\widehat{\tau}^*_{dr,b} - \widehat{\tau}^*_{sz,b}\right)^2 / \hat{V}^*_b.$$

4. Repeat Steps 1-2 B times.

5. Calculate the bootstrapped p-value, p^*, as the proportion of the bootstrap test statistics, $\{\mathscr{T}^*_b\}_{b=1}^B$, that are greater than or equal to $\mathscr{T}_n$.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Abadie, A. (2005). Semiparametric difference-in-difference estimators. *Review of Economic Studies*, 72(1):1–19.

Abbring, J. H. and van den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5):1491–1517.

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic efficiency of semiparametric two-step GMM. *The Review of Economic Studies*, 81(3):919–943,.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418):328–336.

Anstrom, K. J. and Tsiatis, A. A. (2001). Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*, 57(4):1207–1218.

Arellano, M. and Meghir, C. (1992). Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. *The Review of Economic Studies*, 59(3):537–559.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Beran, R. (1981). Nonparametric regression with randomly censored survival data.

Bernasconi, D. P., Antolini, L., et al. (2022). A causal inference approach to compare leukaemia treatment outcome in the absence of randomization and with dependent censoring. *International Journal of Epidemiology*, 51(1):314–323.

Beyhum, J., Florens, J.-P., and Van Keilegom, I. (2021). A nonparametric instrumental approach to endogeneity in competing risks models. *arXiv preprint arXiv:2105.00946*.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, 20(1):105–134.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669.

Bouaziz, O. and Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli*, 16(2):514–542.

Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics*, 33(3):429–447.

Bücher, A., El Ghouch, A., and Van Keilegom, I. (2021). Single-index quantile regression models for censored data. In *Advances in Contemporary Statistics and Econometrics*, pages 177–196. Springer.

Buchinsky, M., Li, F., and Liao, Z. (2022). Estimation and inference of semiparametric models using data from several sources. *Journal of Econometrics*, 226(1):80–103.

Busso, M., Dinardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics*, 96(5):885–895.

Callaway, B., Li, T., and Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Canay, I. A., Santos, A., and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191.

Chen, J. and Roth, J. (2023). Log-like? identified ates defined with zero-valued outcomes are (arbitrarily) scale-dependent. *Working Paper*.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808–843.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):235–251.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013a). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013b). Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737.

Chiang, C.-T. and Huang, M.-Y. (2012). New estimation and inference procedures for a single-index conditional distribution model. *Journal of Multivariate Analysis*, 111:271–285.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 31(6):1852–1884.

Cragg, J. G. and Donald, S. G. (1996). On the asymptotic properties of ldu-based tests of the rank of a matrix. *Journal of the American Statistical Association*, 91(435):1301–1309.

Crommen, G., Beyhum, J., and Van Keilegom, I. (2022). A gaussian model for survival data subject to dependent censoring and confounding. *arXiv preprint arXiv:2208.04184*.

Czado, C. and Van Keilegom, I. (2021). Dependent censoring based on copulas. *arXiv preprint arXiv:2104.06872*.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics*, pages 1157–1167.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15.

Delecroix, M., Hristache, M., and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3):730–769.

Deresa, N. W. and Van Keilegom, I. (2020). Flexible parametric model for survival data subject to dependent censoring. *Biometrical Journal*, 62(1):136–156.

Deresa, N. W., Van Keilegom, I., and Antonio, K. (2022). Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance: Mathematics and Economics*, 107:1–21.

Domínguez, M. A. and Lobato, I. N. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72(5):1601–1615.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403.

Escanciano, J. C. (2006a). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051.

Escanciano, J. C. (2006b). Goodness-of-fit tests for linear and nonlinear time series models. *Journal of the American Statistical Association*, 101(474):531–541.

Fan, J., Heckman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150.

Fan, Y. and Guerre, E. (2016). Multivariate local polynomial estimators: Uniform boundary properties and asymptotic linear representation. In *Essays in Honor of Aman Ullah*. Emerald Group Publishing Limited.

Fan, Y. and Liu, R. (2018). Partial identification and inference in censored quantile regression. *Journal of Econometrics*, 206(1):1–38.

Fan, Y., Sherman, R., and Shum, M. (2014). Identifying treatment effects under data combination. *Econometrica*, 82(2):811–822.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

Firpo, S., Fortin, N., and Lemieux, T. (2009a). Supplement to ‘unconditional quantile regressions’. *Econometrica Supplemental Material*, 77.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009b). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Firpo, S. P., Fortin, N. M., and Lemieux, T. (2018). Decomposing wage distributions using recentered influence function regressions. *Econometrics*, 6(2):28.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pages 1–102. Elsevier, Amsterdam.

Frandsen, B. R. (2015). Treatment effects with censoring and endogeneity. *Journal of the American Statistical Association*, 110(512):1745–1752.

Frölich, M. (2006). Non-parametric regression for binary dependent variables. *The Econometrics Journal*, 9(3):511–540.

Frölich, M. and Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357.

Gatta, G., Capocaccia, R., et al. (2005). Childhood cancer survival trends in europe: a eurocare working group study. *Journal of Clinical Oncology*, 23(16):3742–3751.

Genest, C. and MacKay, J. (1986a). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

Genest, C. and MacKay, R. J. (1986b). Copules archimédiennes et families de lois bidimensionnelles dont les marges sont données. *Canadian journal of statistics*, 14(2):145–159.

Giné, E. and Mason, D. M. (2007). On local u-statistic processes and the estimation of densities of functions of several sample variables. *The Annals of Statistics*, 35(3):1105–1145.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, 22(3):361–411.

Graham, B. S., Pinto, C. C. d. X., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business & Economic Statistics*, 34(2):288–301.

Guerre, E., Perrigne, I., and Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3):525–574.

Guggenberger, P. (2010a). The impact of a hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory*, 26(2):369–382.

Guggenberger, P. (2010b). The impact of a hausman pretest on the size of a hypothesis test: The panel data case. *Journal of Econometrics*, 156(2):337–343.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hájek, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. In Godambe, V. P. and Sprott, D. A., editors, *Foundations of Statistical Inference*. Holt, Rinehart, and Winston, Toronto.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, pages 1251–1271.

Heckman, J. J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Heckman, J. J., Lochner, L. J., and Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. *Handbook of the Economics of Education*, 1:307–458.

Heckman, J. J. and Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hirukawa, M., Murtazashvili, I., and Prokhorov, A. (2020). Yet another look at the omitted variable bias. Working Paper.

Hoeffding, W. et al. (1977). Some incomplete and boundedly complete families of distributions. *The Annals of Statistics*, 5(2):278–291.

Hong, S.-H. (2013). Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes. *Journal of Applied Econometrics*, 28(2):297–324.

Howard, S. C. and Wilimas, J. A. (2005). Delays in diagnosis and treatment of childhood cancer: where in the world are they important? *Pediatric blood & cancer*, 44(4):303–304.

Howlader, N., Noone, A., et al. (2016). Seer cancer statistics review, 1975–2013. bethesda, md: National cancer institute; 2016.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4):1090–1099.

Hubbard, A. E., Laan, M. J., and Robins, J. M. (2000). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 135–177. Springer.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, pages 467–475.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC press.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):569–573.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4):1229–1245.

Khan, S., Ponomareva, M., and Tamer, E. (2016). Identification of panel data models with endogenous censoring. *Journal of Econometrics*, 194(1):57–75.

Khan, S. and Tamer, E. (2009). Inference on endogenously censored regression models using conditional moment inequalities. *Journal of Econometrics*, 152(2):104–119.

Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.

Kim, D. (2021). Partially identifying competing risks models: An application to the war on cancer. *Journal of Econometrics*, 234(2):536–564.

Klevmarken, A. (1982). Missing variables and two-stage least-squares estimation from more than one data set. In *1981 Proceedings of the American Statistical Association, Business and Economic Statistics Section*.

Kline, P. and Santos, A. (2012). A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):1–40.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kong, E., Linton, O., and Xia, Y. (2010). Uniform bahadur representation for local polynomial estimates of m-regression and its application to the additive model. *Econometric Theory*, 26(5):1529–1564.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis*, 84(2):299–318.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer, New York.

Lavergne, P. and Patilea, V. (2013). Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. *Journal of Econometrics*, 177(1):47–59.

Lee, M.-j. and Lee, S.-j. (2005). Analysis of job-training effects on korean women. *Journal of Applied Econometrics*, 20(4):549–562.

Lee, Y. Y. (2018). Efficient propensity score regression estimators of multivalued treatment effects for the treated. *Journal of Econometrics*, 204(2):207–222.

Lehmann, E. (1986). *Testing Statistical Hypotheses (Second Ed.).* Wiley, New York.

Lemieux, T. (2006). The "mincer equation" thirty years after schooling, experience, and earnings. In *Jacob Mincer a pioneer of modern labor economics*, pages 127–145. Springer.

Li, Q. and Ouyang, D. (2005). Uniform convergence rate of kernel estimation with mixed categorical and continuous data. *Economics Letters*, 86(2):291–296.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press, Princeton, New Jersey.

Li, T., Perrigne, I., and Vuong, Q. (2002). Structural estimation of the affiliated private value auction model. *RAND Journal of Economics*, 33(2):171–193.

Li, W. and Patilea, V. (2018). A dimension reduction approach for conditional kaplan–meier estimators. *Test*, 27(2):295–315.

Lo, S. M. and Wilke, R. A. (2010). A copula model for dependent competing risks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):359–376.

Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Communications in Statistics-Theory and Methods*, 40(15):2639–2660.

Lopez, O., Patilea, V., and Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli*, 19(3):721–747.

Maistre, S. and Patilea, V. (2019). Nonparametric model checks of single-index assumptions. *Statistica Sinica*, 29(1):113–138.

Malani, A. and Reif, J. (2015). Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. *Journal of Public Economics*, 124:1–17.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, 21(1):255–285.

Marjerrison, S., Antillon, F., et al. (2013). Outcome of children treated for relapsed acute lymphoblastic leukemia in central america. *Cancer*, 119(6):1277–1283.

Martínez-Iriarte, J. (2023). Sensitivity analysis in unconditional quantile effects.

Martinez-Iriarte, J. and Sun, Y. (2020). Identification and estimation of unconditional policy effects of an endogenous binary treatment. *arXiv preprint arXiv:2010.15864*.

Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica*, 71(5):1339–1375.

Matzkin, R. L. (2007). Nonparametric identification. In *Handbook of Econometrics*, volume 6, pages 5307–5368. Elsevier, Amsterdam.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, 13(2):151–161.

Millimet, D. L. and Tchernis, R. (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business & Economic Statistics*, 27(3):397–415.

Mincer, J. A. et al. (1974). Schooling, experience, and earnings. *NBER Books*.

Mostert, S., Arora, R. S., et al. (2011). Abandonment of treatment for childhood cancer: position statement of a siop podc working group. *The lancet oncology*, 12(8):719–720.

Navarrete, M., Rossi, E., et al. (2014). Treatment of childhood acute lymphoblastic leukemia in central america: A lower-middle income countries experience. *Pediatric blood & cancer*, 61(5):803–809.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Nie, X., Lu, C., and Wager, S. (2019). Nonparametric heterogeneous treatment effect estimation in repeated cross sectional designs. *arXiv preprint arXiv:1905.11622*.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057.

Peterson, A. V. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, 73(1):11–13.

Pollard, D. (1990). Empirical processes: Theory and applications. Ims.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430.

Powell, J. L. and Stoker, T. M. (1996). Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, 75(2):291–316.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130.

Regan, T. L. and Oaxaca, R. L. (2009). Work experience as a source of specification error in earnings models: Implications for gender wage decompositions. *Journal of Population Economics*, 22(2):463–499.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. In *Handbook of Econometrics*, volume 6, pages 5469–5547. Elsevier, Amsterdam.

Rivest, L.-P. and Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1):138–155.

Robin, J.-M. and Smith, R. J. (2000). Tests of rank. *Econometric Theory*, 16(2):151–175.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155(1):56–70.

Rothe, C. (2012). Partial distributional policy effects. *Econometrica*, 80(5):2269–2301.

Rothe, C. and Firpo, S. (2019). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory*, 35(5):1048–1087.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sant'Anna, P. H. (2016). Program evaluation with right-censored data. *arXiv preprint arXiv:1604.02642*.

Sant'Anna, P. H., Song, X., and Xu, Q. (2022). Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6):1093–1120.

Sant'Anna, P. H. (2021). Nonparametric tests for treatment effect heterogeneity with duration outcomes. *Journal of Business & Economic Statistics*, 39(3):816–832.

Sant'Anna, P. H. C. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Sasaki, Y., Ura, T., and Zhang, Y. (2022). Unconditional quantile regression with high-dimensional data. *Quantitative Economics*, 13(3):955–978.

Sequeira, S. (2016). Corruption, trade costs, and gains from tariff liberalization: Evidence from southern africa. *American Economic Review*, 106(10):3029–63.

Sequeira, S. and Djankov, S. (2014). Corruption and firm behavior: Evidence from african ports. *Journal of International Economics*, 94(2):277–294.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Sherman, R. P. (1994). Maximal inequalities for degenerate *u*-processes with applications to optimization estimators. *The Annals of Statistics*, 22(1):439–459.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Smucler, E., Rotnitzky, A., and Robins, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv:1904.03737*.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283.

Strzalkowska-Kominiak, E. and Cao, R. (2014). Beran-based approach for single-index models under censoring. *Computational Statistics*, 29(5):1243–1261.

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., and Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology*, 14(4):166–182.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, pages 613–641.

Tanabe, K. and Sagae, M. (1992). An exact cholesky decomposition and the generalized inverse of the variance–covariance matrix of the multinomial distribution, with applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):211–219.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Univ. Press, Cambridge.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wellner, J. A. W. and van der Vaart, A. W. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: particles, processes and inverse problems*, pages 234–252. Institute of Mathematical Statistics.

Westling, T., Luedtke, A., Gilbert, P., and Carone, M. (2021). Inference for treatment-specific survival curves using machine learning. *arXiv preprint arXiv:2106.06602*.

Xia, Y., Zhang, D., and Xu, J. (2010). Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association*, 105(489):278–290.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138.